



## Amphioxus functional genomics and the origins of vertebrate gene regulation

Ferdinand Marletaz, Panos N. Firbas, Ignacio Maeso, Juan J. Tena, Ozren Bogdanovic, Malcolm Perry, Christopher D. R. Wyatt, Elisa de La Calle-Mustienes, Stephanie Bertrand, Demian Burguera, et al.

### ► To cite this version:

Ferdinand Marletaz, Panos N. Firbas, Ignacio Maeso, Juan J. Tena, Ozren Bogdanovic, et al.. Amphioxus functional genomics and the origins of vertebrate gene regulation. *Nature*, 2018, 564 (7734), pp.64-70. 10.1038/s41586-018-0734-6 . hal-02000511

**HAL Id: hal-02000511**

**<https://hal.science/hal-02000511>**

Submitted on 17 Apr 2019

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution| 4.0 International License

# Amphioxus functional genomics and the origins of vertebrate gene regulation

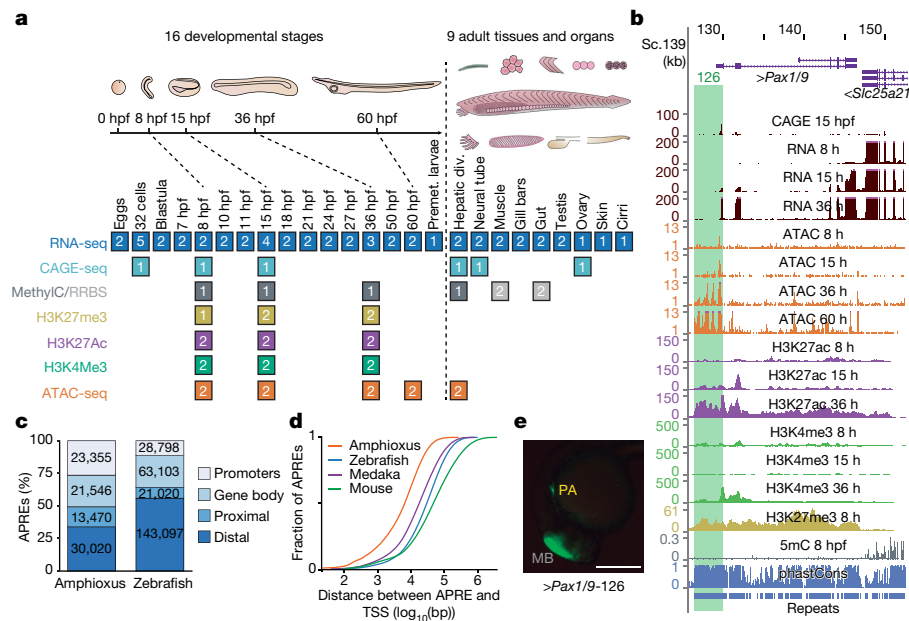
Ferdinand Marlétaz<sup>1,2,41</sup>, Panos N. Firas<sup>3,41</sup>, Ignacio Maeso<sup>3,41\*</sup>, Juan J. Tena<sup>3,41</sup>, Ozren Bogdanovic<sup>4,5,6,41</sup>, Malcolm Perry<sup>7,8,41</sup>, Christopher D. R. Wyatt<sup>9,10</sup>, Elisa de la Calle-Mustienes<sup>3</sup>, Stephanie Bertrand<sup>11</sup>, Demian Burguera<sup>9,12</sup>, Rafael D. Acemel<sup>3</sup>, Simon J. van Heeringen<sup>13</sup>, Silvia Naranjo<sup>3</sup>, Carlos Herrera-Ubeda<sup>12</sup>, Ksenia Skvortsova<sup>4</sup>, Sandra Jimenez-Gancedo<sup>3</sup>, Daniel Aldea<sup>11</sup>, Yamile Marquez<sup>9</sup>, Lorena Buono<sup>3</sup>, Iryna Kozmikova<sup>14</sup>, Jon Permanyer<sup>9</sup>, Alexandra Louis<sup>15,16,17</sup>, Beatriz Albuixech-Crespo<sup>12</sup>, Yann Le Petillon<sup>11</sup>, Anthony Leon<sup>11</sup>, Lucie Subirana<sup>11</sup>, Piotr J. Balwierz<sup>7,8</sup>, Paul Edward Duckett<sup>4</sup>, Ensieh Farahani<sup>3</sup>, Jean-Marc Aury<sup>18</sup>, Sophie Mangelot<sup>18</sup>, Patrick Wincker<sup>19</sup>, Ricard Albalat<sup>20</sup>, Èlia Benito-Gutiérrez<sup>21</sup>, Cristian Cañestro<sup>20</sup>, Filipe Castro<sup>22</sup>, Salvatore D'Aniello<sup>23</sup>, David E. K. Ferrier<sup>24</sup>, Shengfeng Huang<sup>25</sup>, Vincent Laudet<sup>11</sup>, Gabriel A. B. Marais<sup>26</sup>, Pierre Pontarotti<sup>27</sup>, Michael Schubert<sup>28</sup>, Hervé Seitz<sup>29</sup>, Ildiko Somorjai<sup>30</sup>, Tokiharu Takahashi<sup>31</sup>, Olivier Mirabeau<sup>32</sup>, Anlong Xu<sup>25,33</sup>, Jr-Kai Yu<sup>34</sup>, Piero Carninci<sup>35,36</sup>, Juan Ramon Martinez-Morales<sup>3</sup>, Hugues Roest Crollius<sup>15,16,17</sup>, Zbynek Kozmik<sup>14</sup>, Matthew T. Weirauch<sup>37,38</sup>, Jordi Garcia-Fernández<sup>12</sup>, Ryan Lister<sup>6,39</sup>, Boris Lenhard<sup>7,8,40</sup>, Peter W. H. Holland<sup>1</sup>, Hector Escriva<sup>11\*</sup>, Jose Luis Gómez-Skarmeta<sup>3\*</sup> & Manuel Irimia<sup>9,10\*</sup>

Vertebrates have greatly elaborated the basic chordate body plan and evolved highly distinctive genomes that have been sculpted by two whole-genome duplications. Here we sequence the genome of the Mediterranean amphioxus (*Branchiostoma lanceolatum*) and characterize DNA methylation, chromatin accessibility, histone modifications and transcriptomes across multiple developmental stages and adult tissues to investigate the evolution of the regulation of the chordate genome. Comparisons with vertebrates identify an intermediate stage in the evolution of differentially methylated enhancers, and a high conservation of gene expression and its *cis*-regulatory logic between amphioxus and vertebrates that occurs maximally at an earlier mid-embryonic phylotypic period. We analyse regulatory evolution after whole-genome duplications, and find that—in vertebrates—over 80% of broadly expressed gene families with multiple paralogues derived from whole-genome duplications have members that restricted their ancestral expression, and underwent specialization rather than subfunctionalization. Counter-intuitively, paralogues that restricted their expression increased the complexity of their regulatory landscapes. These data pave the way for a better understanding of the regulatory principles that underlie key vertebrate innovations.

All vertebrates share multiple morphological and genomic novelties<sup>1</sup>. The most prominent genomic difference between vertebrates and non-vertebrate chordates is the reshaping of the gene complement that followed the two rounds of whole genome duplication (WGD)—the 2R hypothesis—that occurred at the base of the vertebrate lineage<sup>2,3</sup>. These large-scale mutational events are hypothesized to have

facilitated the evolution of vertebrate morphological innovations, at least in part through the preferential retention of 'developmental' gene families and transcription factors after duplication<sup>3,4</sup>. However, duplicate genes and their associated regulatory elements were initially identical and could not drive innovation without regulatory and/or protein-coding changes.

<sup>1</sup>Department of Zoology, University of Oxford, Oxford, UK. <sup>2</sup>Molecular Genetics Unit, Okinawa Institute of Science and Technology Graduate University, Onna-son, Japan. <sup>3</sup>Centro Andaluz de Biología del Desarrollo (CABD), CSIC-Universidad Pablo de Olavide-Junta de Andalucía, Seville, Spain. <sup>4</sup>Genomics and Epigenetics Division, Garvan Institute of Medical Research, Sydney, New South Wales, Australia. <sup>5</sup>St Vincent's Clinical School, Faculty of Medicine, University of New South Wales, Sydney, New South Wales, Australia. <sup>6</sup>Australian Research Council Centre of Excellence in Plant Energy Biology, School of Molecular Sciences, The University of Western Australia, Crawley, Western Australia, Australia. <sup>7</sup>Institute of Clinical Sciences, Faculty of Medicine, Imperial College London, London, UK. <sup>8</sup>Computational Regulatory Genomics, MRC London Institute of Medical Sciences, London, UK. <sup>9</sup>Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Barcelona, Spain. <sup>10</sup>Universitat Pompeu Fabra (UPF), Barcelona, Spain. <sup>11</sup>Biologie Intégrative des Organismes Marins, BIOM, Observatoire Océanologique, CNRS and Sorbonne Université, Banyuls sur Mer, France. <sup>12</sup>Department of Genetics, Microbiology and Statistics, Faculty of Biology, and Institut de Biomedicina (IBUB), University of Barcelona, Barcelona, Spain. <sup>13</sup>Department of Molecular Developmental Biology, Faculty of Science, Radboud Institute for Molecular Life Sciences, Radboud University, Nijmegen, The Netherlands. <sup>14</sup>Institute of Molecular Genetics of the Czech Academy of Sciences, Prague, Czech Republic. <sup>15</sup>Institut de Biologie de l'ENS, IBENS, Ecole Normale Supérieure, Paris, France. <sup>16</sup>Inserm, U1024, Paris, France. <sup>17</sup>CNRS, UMR 8197, Paris, France. <sup>18</sup>Genoscope, Institut de biologie François-Jacob, Commissariat à l'Energie Atomique (CEA), Université Paris-Saclay, Evry, France. <sup>19</sup>Génomique Métabolique, Genoscope, Institut de biologie François Jacob, Commissariat à l'Energie Atomique (CEA), CNRS, Université Evry, Université Paris-Saclay, Evry, France. <sup>20</sup>Department of Genetics, Microbiology and Statistics, Faculty of Biology and Institut de Recerca de la Biodiversitat (IRBio), University of Barcelona, Barcelona, Spain. <sup>21</sup>Department of Zoology, University of Cambridge, Cambridge, UK. <sup>22</sup>Interdisciplinary Centre of Marine and Environmental Research (CIIMAR/CIMAR) and Faculty of Sciences (FCUP), Department of Biology, University of Porto, Porto, Portugal. <sup>23</sup>Biology and Evolution of Marine Organisms, Stazione Zoologica Anton Dohrn Napoli, Naples, Italy. <sup>24</sup>The Scottish Oceans Institute, Gatty Marine Laboratory, University of St Andrews, St Andrews, UK. <sup>25</sup>State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-sen University, Guangzhou, China. <sup>26</sup>Laboratoire de Biométrie et Biologie Evolutive (UMR 5558), CNRS and Université Lyon 1, Villeurbanne, France. <sup>27</sup>IRD, APHM, Microbe, Evolution, Phyllogénie, Infection, IHU Méditerranée Infection and CNRS, Aix Marseille University, Marseille, France. <sup>28</sup>Sorbonne Université, CNRS, Laboratoire de Biologie du Développement de Villefranche-sur-Mer, Institut de la Mer de Villefranche-sur-Mer, Villefranche-sur-Mer, France. <sup>29</sup>UMR 9002 CNRS, Institut de Génétique Humaine, Université de Montpellier, Montpellier, France. <sup>30</sup>Biomedical Sciences Research Complex, School of Biology, University of St Andrews, St Andrews, UK. <sup>31</sup>School of Medical Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Manchester, UK. <sup>32</sup>INSERM U830, Équipe Labellisée LNCC, SIREDO Oncology Centre, Institut Curie, PSL Research University, Paris, France. <sup>33</sup>School of Life Sciences, Beijing University of Chinese Medicine, Beijing, China. <sup>34</sup>Institute of Cellular and Organismic Biology, Academia Sinica, Taipei, Taiwan. <sup>35</sup>RIKEN Center for Life Science Technologies (Division of Genomic Technologies) (CLST DGT), Yokohama, Japan. <sup>36</sup>Laboratory for Transcriptome Technology, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan. <sup>37</sup>Center for Autoimmune Genomics and Etiology, Divisions of Biomedical Informatics and Developmental Biology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA. <sup>38</sup>Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, OH, USA. <sup>39</sup>Harry Perkins Institute of Medical Research, Nedlands, Western Australia, Australia. <sup>40</sup>Sars International Centre for Marine Molecular Biology, University of Bergen, Bergen, Norway. <sup>41</sup>These authors contributed equally: Ferdinand Marlétaz, Panos N. Firas, Ignacio Maeso, Juan J. Tena, Ozren Bogdanovic, Malcolm Perry. \*e-mail: nacho.maeso@gmail.com; hescriva@obs-banyuls.fr; jlgomska@upo.es; mirimia@gmail.com



**Fig. 1 | Functional genome annotation of amphioxus.** **a**, Summary of the 94 amphioxus samples generated in this study, comprising eight functional-genomic datasets. The number of biological replicates is indicated for each sample type. div., diverticulum; MethyC/RRBS, methylC sequencing and reduced representation bisulfite sequencing; Premet., premetamorphic. **b**, Genome browser excerpt showing a selection of available tracks, including gene annotation, sequence conservation (using phastCons), repeats and several epigenomic and transcriptomic datasets. Green rectangle highlights the APRE tested in **e**. **c**, Numbers and proportions of amphioxus and zebrafish APREs according to their

genomic location. Promoters, within 1-kbp upstream and 0.5-kbp downstream of an annotated TSS; gene body, within an orthology-supported gene; proximal, within 5-kbp upstream of (but not overlapping with) a TSS; distal, not in the aforementioned categories. **d**, Cumulative distributions of the distance between each APRE and the closest annotated TSS in each species. **e**, Lateral view of a representative transgenic zebrafish 26-hpf embryo showing GFP expression driven by an amphioxus APRE associated with *Pax1/9* (*Pax1/9-126*, highlighted in **b**) in pharyngeal arches (PA;  $n = 4/4$ ). Positive-control enhancer was expressed in the midbrain (MB). Scale bar, 250  $\mu$ m.

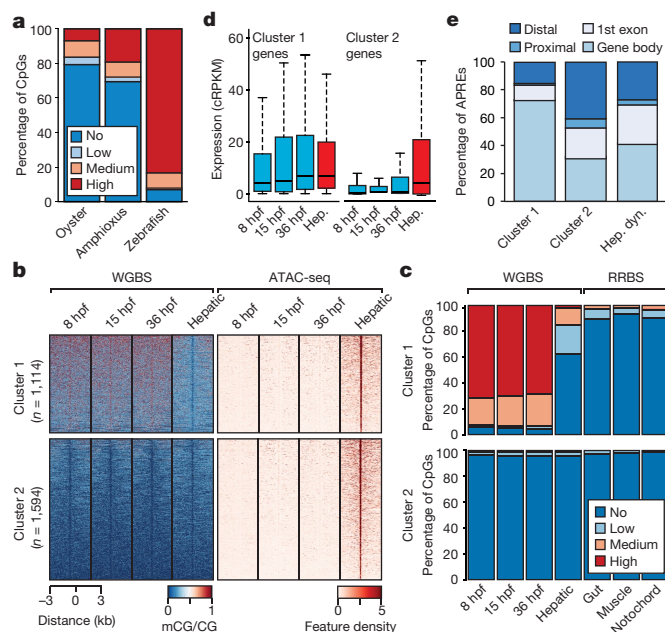
To date, the effect of vertebrate WGDs on gene regulation have remained poorly understood—both in terms of the fates of duplicate genes and the acquisition of the unique genomic traits that are characteristic of vertebrates. These traits include numerous features that are often associated with gene regulation, such as unusually large intergenic and intronic regions<sup>5,6</sup>, high global 5-methylcytosine (5mC) content and 5mC-dependent regulation of embryonic transcriptional enhancers<sup>7</sup>. To investigate these traits, appropriate species must be used for comparisons. Previous studies have largely focused on phylogenetic distances that are either too short (such as human versus mouse) or too long (such as human versus fly or nematode), resulting in limited insights. In the first case, comparisons among closely related species (for example, between mammals<sup>8–11</sup>)—for which the orthology of non-coding regions can be readily determined from genomic alignments—have allowed fine-grained analyses of the evolution of transcription-factor binding. In the second case, three-way comparisons of human, fly and nematode by the modENCODE consortium revealed no detectable conservation at the *cis*-regulatory level<sup>12</sup> and very little conservation of gene expression<sup>13</sup>. Moreover, the genomes of flies and nematodes are highly derived<sup>14–16</sup>. Thus, we lack comprehensive functional genomic data from a slow-evolving, closely related outgroup that would enable an in-depth investigation of the origins of the vertebrate regulatory genome and of the effect of WGDs on gene regulation.

Unlike flies, nematodes and most non-vertebrates, amphioxus belongs to the chordate phylum. Therefore, although amphioxus lacks the specializations and innovations of vertebrates, it shares with them a basic body plan and has multiple organs and structures homologous to those of vertebrates<sup>1</sup>. For these reasons, amphioxus has widely been used as a reference outgroup to infer ancestral versus novel features during vertebrate evolution. Here, we undertook a comprehensive study of the transcriptome and regulatory genome of amphioxus to investigate how the unique functional genome architecture of vertebrates evolved.

## Functional genome annotation of amphioxus

We generated an exhaustive resource of genomic, epigenomic and transcriptomic data for the Mediterranean amphioxus (*B. lanceolatum*), comprising a total of 52 sample types (Fig. 1a and Supplementary Data 2, datasets 1–5). These datasets were mapped to a *B. lanceolatum* genome that was sequenced and assembled de novo, with 150 $\times$  coverage, a total size of 495.4 Mbp, a scaffold N50 of 1.29 Mbp and 4% gaps (Extended Data Fig. 1a–c). To facilitate access by the research community, we integrated these resources into a UCSC Genome Browser track hub (Fig. 1b; available at <http://amphiencode.github.io/Data/>), together with an intra-cephalochordate sequence conservation track and a comprehensive annotation of repetitive elements (Extended Data Fig. 1d–f) and long non-coding RNAs (Extended Data Fig. 1g and Supplementary Data 2, dataset 6). To enable broader evolutionary comparisons, we reconstructed orthologous gene families for multiple vertebrate and non-vertebrate species (Supplementary Data 2, dataset 7), generated several equivalent datasets for zebrafish and medaka (Extended Data Fig. 2a), and built a dedicated server for synteny comparisons (Extended Data Fig. 1h).

A comprehensive functional annotation of the *B. lanceolatum* genome identified 88,391 putative *cis*-regulatory elements of DNA as defined by assay for transposase-accessible chromatin using sequencing (ATAC-seq) (these elements are hereafter referred to as APREs), as well as 20,569 protein-coding genes supported by orthology. We divided the APREs into promoters—around transcription start sites (TSSs), which were highly supported by cap analysis gene-expression sequencing (CAGE-seq) data, Extended Data Fig. 2b—and gene-body, proximal and distal APREs (Fig. 1c). Equivalent analyses using zebrafish data yielded 256,018 potential regulatory regions, with a significantly higher proportion of these being distal APREs (Fig. 1c;  $P < 2.2 \times 10^{-16}$ , one-sided Fisher's exact test). A significantly larger global TSS distance in APREs was observed for all vertebrates compared to amphioxus (Fig. 1d), even after correcting for differences in average intergenic



**Fig. 2 | 5mC patterns and dynamics in the amphioxus genome.** **a**, Percentage of methylated CpG dinucleotides in oyster (mantle,  $n = 14,779,123$ ), amphioxus (8 hpf,  $n = 19,657,388$ ) and zebrafish (1,000-cell stage,  $n = 38,989,847$ ) samples. Low,  $>0$ –20%; medium, 20–80%; high,  $>80\%$ . **b**,  $k$ -means clustering ( $n = 2$ ) of 5mC signal over hepatic-specific APREs. **c**, Percentage of methylated CpG dinucleotides as assessed by whole-genome bisulfite sequencing (WGBS) and reduced representation bisulfite sequencing (RRBS) in embryos and adult tissues in APREs from **b**. **d**, Distribution of expression levels for genes associated with APREs displaying distinct 5mC patterns in **b**. Cluster 1: 1,114 genes; cluster 2: 1,594 genes. cRPKM, corrected (per mappability) reads per kb of mappable positions and million reads. Hep, hepatic diverticulum. **e**, Genomic distribution of regions with distinct 5mC patterns from **b**. Hep. dyn., dynamic APREs active in the hepatic diverticulum.

length among species (Extended Data Fig. 2c;  $P < 2.2 \times 10^{-16}$  for all vertebrate-versus-amphioxus comparisons, one-sided Mann–Whitney tests). Amphioxus APREs showed enrichment for enhancer-associated

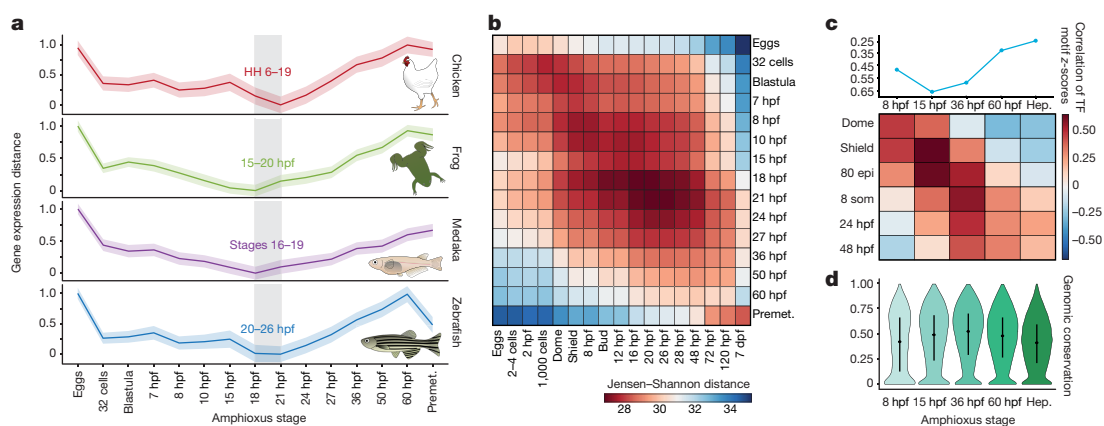
chromatin marks (Extended Data Fig. 2d), which were highly dynamic during embryo development (Extended Data Fig. 2e–g), and consistently drove GFP expression in zebrafish or amphioxus transgenic assays (93% (14/15), Fig. 1e and Extended Data Fig. 2h, i). Moreover, 89% (32/36) of previously reported amphioxus enhancers overlapped APREs defined by our data. Therefore, a large fraction of APREs probably act as developmentally regulated transcriptional enhancers.

### Disentangling vertebrate bidirectional promoters

Analyses of core promoters, defined by CAGE-seq, at single-nucleotide resolution revealed that amphioxus promoters display a mixture of pan-metazoan, pan-vertebrate and unique features (Extended Data Fig. 3 and Supplementary Information). These analyses also identified that 25% (3,950/15,884) of neighbouring protein-coding genes were arranged in bidirectional promoters. Bidirectional promoters were most common among ubiquitous promoters (Extended Data Fig. 4a), displayed a marked periodicity in the distance between promoters (Extended Data Fig. 4b, c) and were associated with genes that were significantly enriched in housekeeping functions (Extended Data Fig. 4d). Notably, the fraction of bidirectional promoters defined by CAGE-seq decreased progressively from amphioxus to mouse (12.83% (1,752/13,654)) and to zebrafish (7.84% (1,098/14,014)), which suggests a disentanglement of ancestral bidirectional promoters after each round of WGD (two in tetrapods and three in teleosts). Consistently, the majority of a set of 372 putatively ancestral, bidirectional promoters were lost in vertebrates—particularly in stem vertebrates (54.5%)—with only very few amphioxus-specific losses (5.3%) (Extended Data Fig. 4e, f).

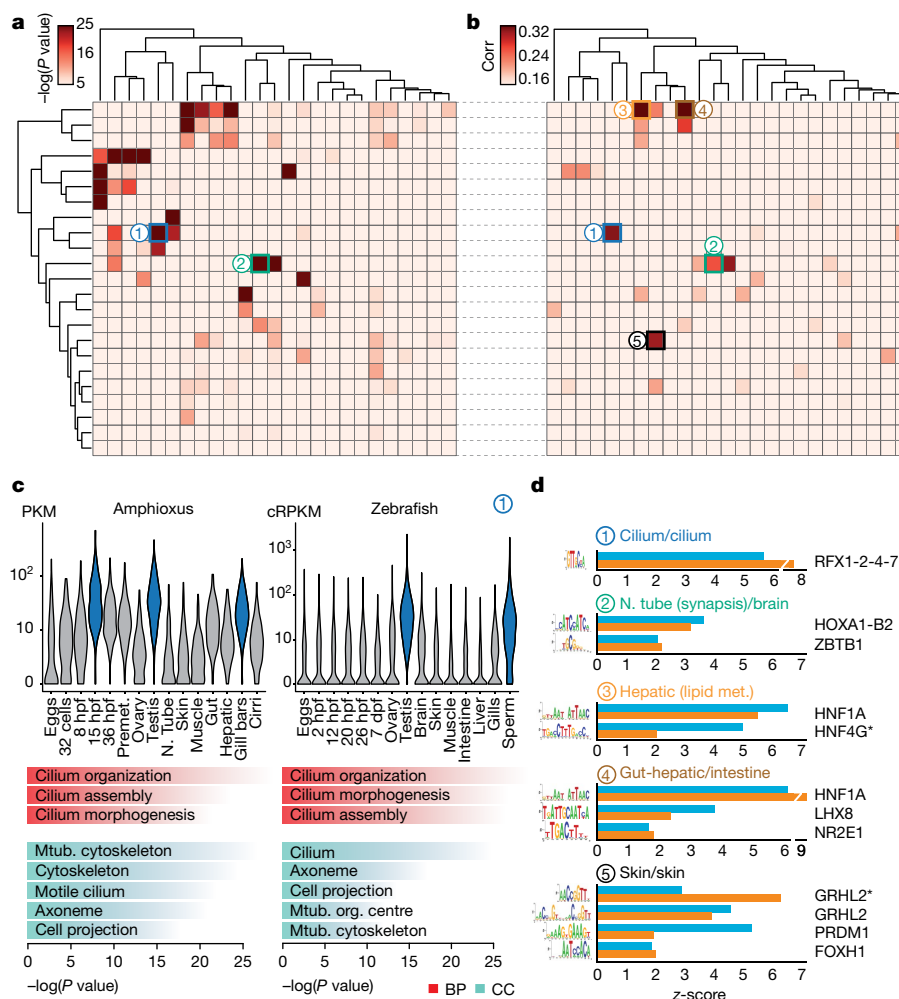
### Developmental DNA demethylation of APREs

Similar to other non-vertebrates<sup>17–19</sup>, the amphioxus genome exhibited very low levels of CpG methylation (Fig. 2a); nearly all of the 5mC occurred in gene bodies, in which the proportion of methylated CpGs correlated positively with gene-expression levels but negatively with the density of H3K27me3 and H3K4me3 histone marks and CpG dinucleotides (Extended Data Fig. 5a–c). However, as in zebrafish and frogs<sup>7</sup>, global levels of 5mC displayed a decrease during development (Extended Data Fig. 5d–g), coinciding with the onset of expression of the amphioxus orthologue of TET demethylase (Extended Data Fig. 5h).



**Fig. 3 | The hourglass model and chordate embryogenesis.** **a**, Stages of minimal transcriptomic divergence (using the Jensen–Shannon distance metric) from four vertebrate species to each amphioxus stage. The grey box outlines the period of minimal divergence, with the corresponding vertebrate periods indicated (the range is given by the two less divergent stages). Dispersions correspond to the standard deviation computed on 100 bootstrap re-samplings of the orthologue sets (amphioxus–chicken: 5,720; amphioxus–zebrafish: 5,673; amphioxus–frog: 5,883; and amphioxus–medaka: 5,288). HH, Hamburger–Hamilton stage. **b**, Heat map of pairwise transcriptomic Jensen–Shannon distances between amphioxus (vertical) and zebrafish (horizontal) stages. A smaller distance (red) indicates higher similarity. dpf, days post-fertilization. **c**, Zebrafish and amphioxus pairwise Pearson correlation of relative enrichment z-scores for transcription-factor (TF) motifs in dynamic APREs, active at different developmental stages. Top, maximal correlation for each amphioxus stage against the zebrafish stages. Bottom, heat map with all pairwise correlations. 80 epi, 80% epiboly stage; 8 som, 8-somite stage. **d**, Sequence conservation levels within the cephalochordates of active APREs at each developmental stage, visualized as the distribution of average phastCons scores. The number of APREs at 8 hpf = 5,282; at 15 hpf = 17,387; at 36 hpf = 21,089; at 60 hpf = 22,674; and in hepatic diverticulum (hep) = 16,551. Dots correspond to the mean values and lines represent the interquartile range.





**Fig. 4 | Transcriptomic and *cis*-regulatory conservation of adult chordate tissues.** **a**, Heat map showing the level of raw statistical significance of orthologous gene overlap between modules produced by weighted correlation network analysis (WGCNA), from amphioxus (vertical) and zebrafish (horizontal) as derived from upper-tail hypergeometric tests. **b**, Heat map of all pairwise Pearson correlations (corr) between the modules of the two species, based on the relative z-scores of transcription-factor motifs for each module (242 super-families of motifs). Modules are clustered as in **a**. **c**, Distribution of expression values (cRPKM) for all genes within the cilium modules across each sample (top), and enriched Gene Ontology terms within each module (bottom) for a pair of modules (labelled '1' in **b**; 1,681 and 261 genes in zebrafish and amphioxus, respectively). BP, biological process; CC, cellular component. *P* values correspond to uncorrected two-sided Fisher's exact tests as provided by topGO. Mtub., microtubule; N. tube, neural tube; org., organizing. **d**, Transcription-factor binding-site motifs with high z-scores from highly correlated pairs of modules between zebrafish (blue) and amphioxus (orange). Numbers correspond to those circles in **b**. RFX1-2-4-7 denotes RFX1, RFX2, RFX4 and RFX7; HOXA1-B2 denotes HOXA1 and HOXB2; asterisk denotes an alternative motif.

To assess whether these 5mC dynamics may have regulatory potential, we identified adult hepatic diverticulum-specific APREs that are inactive during development. Unlike embryo-specific APREs (Extended Data Fig. 6a), the clustering of these adult APREs on the basis of 5mC content revealed two distinct subsets, one with hepatic-specific and one with constitutive hypomethylation (Fig. 2b). Differentially methylated APREs (cluster 1) also displayed robust hypomethylation in other adult tissues (Fig. 2c), which suggests that demethylation at these APREs occurs organism-wide. Both groups of hepatic-specific APREs were enriched for binding sites of liver-specific transcription factors—such as Hnf4a—as well as broadly expressed transcription factors such as Foxa (Extended Data Fig. 6b), which is a pioneer factor that participates in 5mC removal at regulatory regions in mammals<sup>20</sup>.

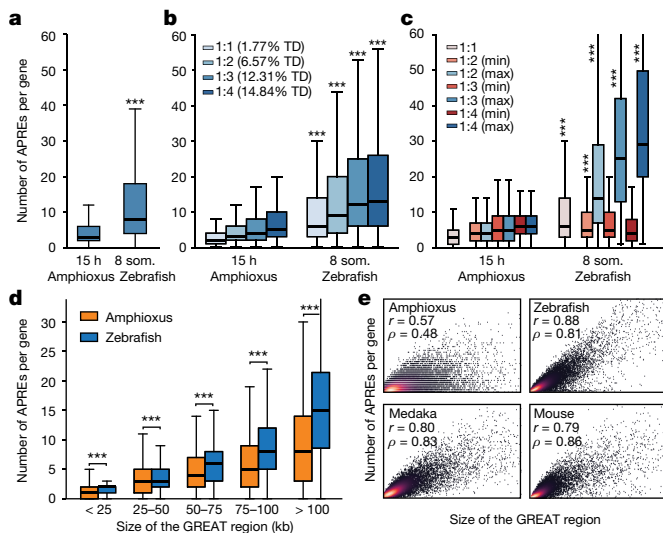
APREs from both clusters were preferentially associated with genes with metabolic functions (Extended Data Fig. 6c). However, only APREs with hepatic-specific hypomethylation (cluster 1) were primarily associated with genes that displayed steady widespread expression (Fig. 2d and Extended Data Fig. 6d, e); these APREs were mainly located within gene bodies (Fig. 2e). These data suggest that demethylation of these APREs may contribute to their identification as adult-specific, transcriptional *cis*-regulatory elements within continuously hypermethylated gene-body contexts, which is characteristic of non-vertebrate species. Fourteen zebrafish gene families contained differentially methylated APREs in introns that are orthologous to those identified in amphioxus—amongst these are four genes that encode components of the Hippo pathway, including the transcriptional effectors Yap (*yap1* and *wvtr1*) and Tead (*tead1a* and *tead3a*) (Extended Data Fig. 6f, g).

## The hourglass model and chordate embryogenesis

Previous comparative analyses among vertebrate transcriptomes<sup>21,22</sup> showed a developmental period of maximal similarity in gene expression that coincides with the so-called phylotypic period, consistent with the hourglass model<sup>23</sup>. However, similar comparisons with tunicates and amphioxus have thus far not resolved a phylotypic period shared across all chordates<sup>22</sup>. Pairwise comparisons of stage-specific RNA sequencing (RNA-seq) data from developmental time courses of amphioxus against zebrafish, medaka, frog (*Xenopus tropicalis*) and chicken revealed a consistent period of highest similarity (Fig. 3a, b and Extended Data Fig. 7) that occurred slightly earlier than those reported for vertebrates; in amphioxus, this corresponds to the neurula at the 4–7-somite stage (18–21 hours post fertilization (hpf)). At the regulatory level, pairwise comparisons between the relative enrichment of transcription-factor motifs in sets of dynamic APREs that were active at each stage were also consistent with an earlier hourglass model<sup>24</sup> (Fig. 3c). By contrast, at a shorter timescale, comparisons between different species of amphioxus showed that the sequence conservation for the same APREs was higher after the putative chordate phylotypic period (Fig. 3d).

## Regulatory conservation shapes chordate body plan

Additional comparisons of embryo transcriptomes and neighbourhood analysis of conserved co-expression<sup>25</sup> showed a high conservation of developmental and global expression patterns and of gene functions between amphioxus and vertebrates (Extended Data Fig. 8 and Supplementary Information). Further pairwise comparison of co-regulated gene modules across tissues between amphioxus and zebrafish revealed multiple pairs with highly significant levels of orthologue overlap (Fig. 4a). These included modules with conserved tissue-specific



**Fig. 5 | Higher regulatory complexity in vertebrate regulatory landscapes.** **a**, Distribution of the number of APREs within the regulatory landscape of each gene (as estimated by GREAT), at comparable pre-phylogenetic developmental stages (15 hpf for amphioxus and 8 somites for zebrafish).  $n = 6,047$  and  $9,239$  genes for amphioxus and zebrafish, respectively. **b**, As in **a**, but with gene families separated according to the number of retained ohnologues per family in vertebrates (from 1 to 4, using mouse as a reference). The percentage of developmental regulatory genes (TD, trans-dev) in each category is indicated. **c**, As in **b**, but only the genes with the lowest ('min', in red) and the highest ('max', in blue) number of APREs are plotted for each ohnologue family. **d**, Distributions of the number of APREs per gene among subsets of amphioxus and zebrafish genes matched by GREAT-region size ( $\pm 500$  bp) and binned by size as indicated. **e**, Density scatter plot of the number of APREs (y axis) versus the size of the GREAT region (x axis) per gene and species. Pearson ( $r$ ) and Spearman ( $\rho$ ) correlation coefficients are indicated. Sample sizes: amphioxus, 20,053; zebrafish, 20,569; medaka, 15,978; mouse, 18,838. **a–d**, \*\*\*  $P < 0.001$ ; one-sided Mann–Whitney tests of the zebrafish distribution versus the equivalent amphioxus distribution. Exact  $P$  values and sample sizes are provided in Supplementary Data 2, dataset 8.

expression that were enriched for coherent Gene Ontology categories, including genes with high expression in organs with ciliated cells (for example, spermatozoa and gill bars) (labelled '1' in Fig. 4a–c) as well as neural, muscle, gut, liver, skin and metabolism-related modules (Supplementary Data 1). We also found a significant positive correlation between relative motif-enrichment scores for many pairs of modules (Fig. 4b); the most-enriched transcription-factor motifs within each cluster were highly consistent between amphioxus and zebrafish (Fig. 4d).

### Higher regulatory information in vertebrate genomes

To investigate the effect of WGDs on the evolution of vertebrate gene regulation, we first asked whether the number of putative regulatory regions per gene is higher in vertebrates than in amphioxus. We observed significantly more APREs in the regulatory landscape of each gene (as defined by the 'Genomic Regions Enrichment of Annotations Tool' (GREAT)<sup>26</sup>) in zebrafish than in amphioxus (Fig. 5a). This difference is particularly evident for gene families that have retained multiple copies after WGD (known as ohnologues; Fig. 5b), for which the number of APREs is very uneven between copies, with marked regulatory expansions observed for some ohnologues (Fig. 5c). The same patterns were detected for all developmental stages of amphioxus and zebrafish, as well as for medaka and mouse genomes, and were highly robust to down-sampling of ATAC-seq coverage in vertebrates (Extended Data Fig. 9a–c). We also detected a higher number of peaks associated with regulatory genes ('trans-dev' genes that are involved in the regulation of embryonic development) compared to housekeeping genes in all species (Extended Data Fig. 9d), consistent

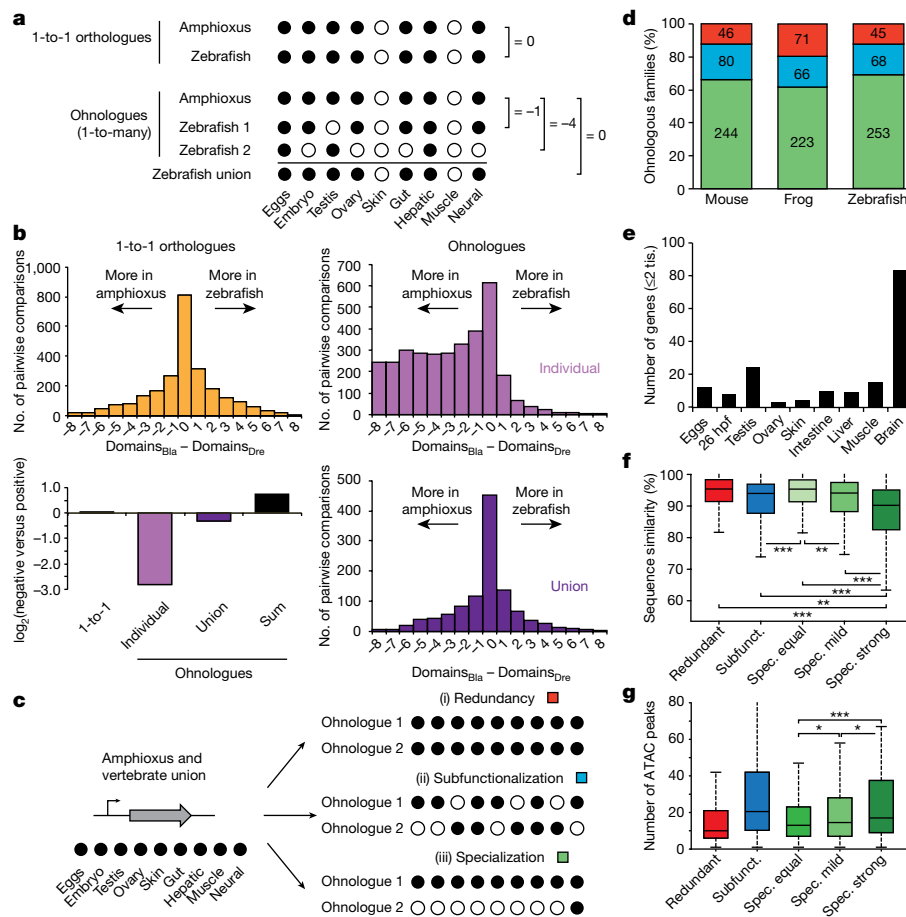
with the higher frequency of retention of trans-dev genes in multiple copies after WGD<sup>3</sup> (Fig. 5b). Comparison of regulatory landscapes—determined experimentally using circular chromosome conformation capture followed by sequencing (4C-seq)—for 58 genes from 11 trans-dev gene families in amphioxus, zebrafish and mouse showed similar results (Extended Data Fig. 9e).

As expected, the higher number of APREs in zebrafish was associated with larger intergenic regions in this species (Extended Data Fig. 9f). However, the differences in APRE complements were not attributable only to an increase in genome size in vertebrates, as subsets of amphioxus and zebrafish genes with matched distributions of GREAT or intergenic-region lengths also displayed a higher number of APREs in zebrafish (Extended Data Fig. 9g, h). Further investigation of matched distributions showed that these differences were particularly great in genes with large regulatory landscapes ( $>50$  kb) (Fig. 5d). Thus, larger regions in amphioxus did not scale at the same rate as in vertebrates in terms of regulatory complexity (Fig. 5e), which is consistent with the overall lower proportion of distal APREs identified in this species (Fig. 1c, d). In summary, these analyses reveal a large increase in the number of regulatory regions during vertebrate evolution (and/or a decrease in these regions in amphioxus)—particularly of distal regulatory elements—and that this trend is enhanced for specific gene copies retained after the WGDs, pointing to unequal rates of regulatory evolution for different ohnologues.

### More-complex regulation in specialized ohnologues

The duplication–degeneration–complementation (DDC) model hypothesizes that the retention of duplicate genes could be driven by reciprocal loss of regulatory elements and restriction of paralogues to distinct subsets of the ancestral expression pattern<sup>27</sup>. In particular, the DDC model predicts that individual paralogues would each have more restricted expression than an unduplicated outgroup, but that their summation would not. To test this, we binarized the expression ('on' or 'off') of each gene in nine homologous expression domains in amphioxus, zebrafish, frog and mouse (Fig. 6a). When comparing genes that returned to single-copy status after WGDs, we detected no expression bias between amphioxus and vertebrates (Fig. 6a, b and Extended Data Fig. 10a, b). By contrast, when vertebrate ohnologues were compared to their single amphioxus orthologues, the distributions were strongly skewed and many vertebrate genes displayed far more restricted expression domains (Fig. 6b and Extended Data Fig. 10a, b; similar results were obtained by comparing  $\tau$  values<sup>28</sup>, Extended Data Fig. 10c–e). The symmetrical pattern was fully recovered when the expression of all vertebrate members was combined, or when the raw expression values were summed for each member within a paralogy group (Fig. 6a, b and Extended Data Fig. 10a, b).

Although the above findings are consistent with the DDC model, they are also compatible with an alternative model in which a subset of duplicate genes becomes more 'specialized' in expression pattern while one or more paralogues retain the broader ancestral expression<sup>29</sup>. To distinguish between these alternatives, we analysed a subset of multi-gene families in which both the single amphioxus orthologue and the union of the vertebrate ohnologues—and thus probably the ancestral gene—were expressed across all nine samples that we compared. We then identified (i) gene families in which all vertebrate paralogues were expressed in all domains (termed 'redundancy'), (ii) gene families in which none of the vertebrate members had expression across all domains (termed 'subfunctionalization')<sup>27</sup> and (iii) gene families in which one or more vertebrate ohnologues were expressed in all domains, but at least one ohnologue was not (termed 'specialization') (Fig. 6c). We obtained very similar results for the three vertebrate species we studied (Fig. 6d): between 80 and 88% of gene families were subfunctionalized or specialized, which implies that ancestral expression domains have been lost in at least one member. Moreover, specialization was consistently more frequent than subfunctionalization as a fate for ohnologues with broad ancestral expression.



**Fig. 6 | Expression specialization is the main fate after WGD.**

**a**, Schematic of the analysis shown in **b**. Expression is binarized for each gene across the nine homologous samples ('on', black dots; normalized cRPKM > 5). **b**, Distribution of the difference in positive domains between zebrafish (domains<sub>Dre</sub>) and amphioxus (domains<sub>Bla</sub>) for 1-to-1 orthologues (2,478 gene pairs; yellow), individual ohnologues (3,427 gene pairs in 1,135 families; lilac) and the union of all vertebrate ohnologues in a family (purple). Bottom left, log<sub>2</sub> of the ratio between zebrafish genes with negative and positive score for each category. 'Sum' (black), binarization of family expression after summing the raw expression values for all ohnologues. **c**, Schematic of the analyses shown in **d**, representing the three possible fates after WGD. **d**, Distribution of fates after WGD for families of ohnologues. **e**, Number of ohnologues with strong

specialization in zebrafish expressed in each domain. Tis., tissue. **f**, Distribution of the percentage of nucleotide sequence similarity between human and mouse by family type. Ohnologues from specialized families are divided into 'spec. equal' (which maintain all expression domains), 'spec. mild' (which have lost some but maintained more than two expression domains) and 'spec. strong' (<=2 expression domains remain). Subfunct., subfunctionalized. **g**, Distribution of the number of APREs within GREAT regions for zebrafish ohnologues for each category. Only statistical comparisons within specialized families are shown. *P* values in **f** and **g** correspond to two- and one-sided Wilcoxon sum-rank tests between the indicated groups, respectively. \*0.05 > *P* value ≥ 0.01, \*\*0.01 > *P* value ≥ 0.001, \*\*\**P* value < 0.001. Exact *P* values and sample sizes are provided in Supplementary Data 2, dataset 8.

Ohnologues that have experienced strong specialization (<=2 remaining expression domains) retained expression more often in neural tissues (Fig. 6e and Extended Data Fig. 10f–i) and were generally not expressed in additional vertebrate-specific tissues (Supplementary Information). Furthermore, they showed the fastest rates of sequence evolution (Fig. 6f and Extended Data Fig. 10j–l), consistent with an optimization of their coding sequence to perform their function in a specific tissue and/or with the evolution of novel functions (neofunctionalization). Ohnologues from specialized families that have lost expression domains showed significantly more associated APREs than ohnologues with the full ancestral expression (Fig. 6g). We observed a strong positive relationship between the number of ancestral expression domains lost and the number of APREs associated with specialized ohnologues (Extended Data Fig. 10m). This implies that the specialization of gene expression after WGD does not occur primarily through loss of ancestral tissue-specific enhancers, but rather by a complex remodelling of regulatory landscapes that involves recruitment of novel, tissue-specific regulatory elements.

## Discussion

By applying functional genomics approaches to the cephalochordate amphioxus, we have deepened our understanding of the origin and

evolution of chordate genomes. We identified APREs in amphioxus, the activation of which is tightly associated with differential DNA demethylation in adult tissues—a mechanism previously thought to be specific to vertebrates. Additional cases may be subsequently found in other non-vertebrate species when similar multi-omics datasets are analysed. In amphioxus, APREs of this type usually fall within gene bodies of widely expressed genes, which suggests that gene regulation by demethylation could have originated as a mechanism to allow better definition of enhancers in a hyper-methylated intra-genic context. If so, this mechanism could have been co-opted into new genomic contexts—that is, distal intergenic enhancers—later in the evolution of vertebrate genomes, which are characterized by their pervasive, genome-wide hypermethylation.

We also found a consistently higher number of open chromatin regions per gene in vertebrates than in amphioxus. This pattern is observed at a genome-wide level, but is particularly evident for distal APREs and in gene families that retain multiple ohnologues after WGD; these families are enriched for regulatory genes with large regulatory landscapes. Finally, we detected a large degree of specialization in expression for retained ohnologues, with the vast majority of multi-gene families with broad ancestral expression having at least one member



that restricted its expression breadth. Through this mechanism, vertebrates have increased their repertoire of tightly regulated genes, which has potentially contributed to tissue-specific evolution. Gene-expression specialization was accompanied by faster evolution of protein-coding sequences, and by an increase—rather than a decrease—in the number of regulatory elements. Taken together, these observations indicate that the two rounds of WGD not only caused an expansion and diversification of gene repertoires in vertebrates, but also allowed functional and expression specialization of the extra copies by increasing the complexity of their gene regulatory landscapes. We suggest that these changes to the gene regulatory landscapes underpinned the evolution of morphological specializations in vertebrates.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0734-6>.

Received: 22 November 2017; Accepted: 18 October 2018;

Published online 21 November 2018.

- Bertrand, S. & Escriva, H. Evolutionary crossroads in developmental biology: amphioxus. *Development* **138**, 4819–4830 (2011).
- Dehal, P. & Boore, J. L. Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol.* **3**, e314 (2005).
- Putnam, N. H. et al. The amphioxus genome and the evolution of the chordate karyotype. *Nature* **453**, 1064–1071 (2008).
- Holland, L. Z. et al. The amphioxus genome illuminates vertebrate origins and cephalochordate biology. *Genome Res.* **18**, 1100–1111 (2008).
- International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
- Nelson, C. E., Hersh, B. M. & Carroll, S. B. The regulatory content of intergenic DNA shapes genome architecture. *Genome Biol.* **5**, R25 (2004).
- Bogdanović, O. et al. Active DNA demethylation at enhancers during the vertebrate phylotypic period. *Nat. Genet.* **48**, 417–426 (2016).
- Berthelot, C., Villar, D., Horvath, J. E., Odom, D. T. & Flicek, P. Complexity and conservation of regulatory landscapes underlie evolutionary resilience of mammalian gene expression. *Nat. Ecol. Evol.* **2**, 152–163 (2018).
- Reilly, S. K. et al. Evolutionary changes in promoter and enhancer activity during human corticogenesis. *Science* **347**, 1155–1159 (2015).
- Villar, D. et al. Enhancer evolution across 20 mammalian species. *Cell* **160**, 554–566 (2015).
- Vierstra, J. et al. Mouse regulatory DNA landscapes reveal global principles of cis-regulatory evolution. *Science* **346**, 1007–1012 (2014).
- Boyle, A. P. et al. Comparative analysis of regulatory information and circuits across distant species. *Nature* **512**, 453–456 (2014).
- Gerstein, M. B. et al. Comparative analysis of the transcriptome across distant species. *Nature* **512**, 445–448 (2014).
- Hendrich, B. & Tweedie, S. The methyl-CpG binding domain and the evolving role of DNA methylation in animals. *Trends Genet.* **19**, 269–277 (2003).
- Irimia, M. et al. Extensive conservation of ancient microsynteny across metazoans due to cis-regulatory constraints. *Genome Res.* **22**, 2356–2367 (2012).
- Simakov, O. et al. Insights into bilaterian evolution from three spiral genomes. *Nature* **493**, 526–531 (2013).
- Wang, X. et al. Genome-wide and single-base resolution DNA methylomes of the Pacific oyster *Crassostrea gigas* provide insight into the evolution of invertebrate CpG methylation. *BMC Genomics* **15**, 1119 (2014).
- Albalat, R., Martí-Solans, J. & Cañestro, C. DNA methylation in amphioxus: from ancestral functions to new roles in vertebrates. *Brief. Funct. Genomics* **11**, 142–155 (2012).
- Huang, S. et al. Decelerated genome evolution in modern vertebrates revealed by analysis of multiple lancelet genomes. *Nat. Commun.* **5**, 5896 (2014).
- Zhang, Y. et al. Nucleation of DNA repair factors by FOXA1 links DNA demethylation to transcriptional pioneering. *Nat. Genet.* **48**, 1003–1013 (2016).
- Irie, N. & Kuratani, S. Comparative transcriptome analysis reveals vertebrate phylotypic period during organogenesis. *Nat. Commun.* **2**, 248 (2011).
- Hu, H. et al. Constrained vertebrate evolution by pleiotropic genes. *Nat. Ecol. Evol.* **1**, 1722–1730 (2017).
- Duboule, D. Temporal colinearity and the phylotypic progression: a basis for the stability of a vertebrate Bauplan and the evolution of morphologies through heterochrony. *Development* **1994 Suppl.**, 135–142 (1994).
- Bogdanović, O. et al. Dynamics of enhancer chromatin signatures mark the transition from pluripotency to cell specification during embryogenesis. *Genome Res.* **22**, 2043–2053 (2012).
- Yue, F. et al. A comparative encyclopedia of DNA elements in the mouse genome. *Nature* **515**, 355–364 (2014).
- McLean, C. Y. et al. GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.* **28**, 495–501 (2010).
- Force, A. et al. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**, 1531–1545 (1999).
- Yanai, I. et al. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* **21**, 650–659 (2005).
- Sandve, S. R., Rohlfs, R. V. & Hvidsten, T. R. Subfunctionalization versus neofunctionalization after whole-genome duplication. *Nat. Genet.* **50**, 908–909 (2018).

**Acknowledgements** This research was funded primarily by the European Research Council (ERC) under the European Union's Horizon 2020 and Seventh Framework Program FP7 research and innovation programs (ERC-AdG-LS8-740041 to J.L.G.-S., ERC-StG-LS2-637591 to M.I., a Marie Skłodowska-Curie Grant (658521) to I.M. and a FP7/2007–2013-ERC-268513 to P.W.H.H.), the Spanish Ministerio de Economía y Competitividad (BFU2016-74961-P to J.L.G.-S., RYC-2016-20089 to I.M., BFU2014-55076-P and BFU2017-89201-P to M.I. and BFU2014-55738-REDT to J.L.G.-S., M.I. and J.R.M.-M.), the 'Centro de Excelencia Severo Ochoa 2013–2017' (SEV-2012-0208), the 'Unidad de Excelencia María de Maetzu 2017–2021' (MDM-2016-0687), the People Program (Marie Curie Actions) of the European Union's Seventh Framework Program FP7 under REA grant agreement number 607142 (DevCom) to J.L.G.-S., and the CNRS and the ANR (ANR16-CE12-0008-01) to H.E. O.B. was supported by an Australian Research Council Discovery Early Career Researcher Award (DECRA; DE140101962). We acknowledge the support of the CERCA Programme/Generalitat de Catalunya and of the Spanish Ministry of Economy, Industry and Competitiveness (MEIC) to the EMBL partnership. Additional sources of funding for all authors are listed in Supplementary Information.

**Reviewer information** *Nature* thanks D. Duboule and P. Flicek for their contribution to the peer review of this work.

**Author contributions** F.M., P.N.F., I.M., J.J.T., O.B., M.P., B.L., P.W.H.H., H.E., J.L.G.-S. and M.I. contributed to concept and study design. F.M., P.N.F., I.M., J.J.T., O.B., M.P., C.D.R.W., R.D.A., S.J.v.H., C.H.-U., K.S., Y.M., A. Louis, P.J.B., P.E.D., M.T.W., J.G.-F., R.L., B.L., P.W.H.H., J.L.G.-S. and M.I. performed computational analyses and data interpretation. O.B., E.d.I.C.-M., S.B., D.B., R.D.A., S.N., S.J.-G., D.A., L.B., J.P., B.A.-C., Y.L.P., A. Leon, L.S., E.F., P.C., J.R.M.-M., R.L., B.L., H.E., J.L.G.-S. and M.I. obtained biological material and generated next-generation sequencing data. I.M., J.J.T., E.d.I.C.-M., I.K., R.D.A., Z.K. and J.L.G.-S. performed transgenic assays. J.-M.A., S.M. and P.W. sequenced the genome. R.A., E.B.-G., C.C., F.C., S.D., D.E.K.F., S.H., V.L., G.A.B.M., P.P., M.S., H.S., I.S., T.T., O.M., A.X. and J.-K.Y. contributed to genome sequencing and gene family curation. I.M., H.E., J.L.G.-S. and M.I. coordinated the project. F.M., I.M., P.W.H.H. and M.I. wrote the main text, with input from all authors. Detailed contributions are listed in Supplementary Information. Animal illustrations by J.J.T., released under a Creative Commons Attribution (CC-BY) Licence.

**Competing interests** The authors declare no competing interests.

## Additional information

**Extended data** is available for this paper at <https://doi.org/10.1038/s41586-018-0734-6>.

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41586-018-0734-6>.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>.

**Correspondence and requests for materials** should be addressed to I.M., H.E., J.L.G. or M.I.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.



## METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment.

**Animal husbandry and embryo staging.** Amphioxus gametes were obtained by heat stimulation as previously described<sup>30,31</sup>. Embryos were obtained by in vitro fertilization in filtered seawater and cultured at 19 °C. Staging was done based on previous publications<sup>32,33</sup>; correspondence between developmental stages and hpf are provided in Supplementary Table 1. All protocols used for vertebrate species (zebrafish and medaka) have been approved by the Institutional Animal Care and Use Ethic Committee (PRBB–IACUEC, for CRG) or the Ethics Committee of the Andalusian Government (license numbers 450–1839 and 182–41106, for CABD–CSIC), and implemented according to national and European regulations. All experiments were carried out in accordance with the principles of the 3Rs (replacement, reduction and refinement).

**Genome sequencing and assembly.** Genomic DNA was extracted from a single *B. lanceolatum* adult male collected in Argeles-sur-Mer, France. The genome was sequenced using a combination of Illumina libraries from a range of inserts at Genoscope (897 million reads in total, with a paired-end coverage of 150×; Supplementary Table 2). A diploid assembly was generated using SOAPdenovo assembler<sup>34</sup> using a k-mer of 71. After gap closing, haplotypes were reconciled with Haplomerger<sup>35</sup>.

**Genome annotation.** We generated deep coverage RNA-seq for 16 developmental stages and 9 adult tissues (4.16 billion reads in total). The bulk of strand-specific transcriptomic data was assembled de novo with Trinity<sup>36</sup>, aligned and assembled into loci with the PASA pipeline<sup>37</sup>. De novo gene models were built using Augustus<sup>38</sup> and subsequently refined with EVM<sup>39</sup> using PASA assemblies and aligned proteins from other species. In parallel, all strand-specific RNA-seq reads were mapped to the genome using Tophat2<sup>40</sup>, assembled using Cufflinks<sup>41</sup> and open reading frames were predicted using Trans-decoder<sup>42</sup>. Models obtained using both these approaches were reconciled yielding a total 218,070 transcripts from 90,927 unified loci, of which 20,569 were protein-coding and had homologues in at least one of the other studied species (see ‘Comparative genomics’). Gene Ontology (GO) terms were assigned to amphioxus proteins based on their PFAM and Interpro domains, as well as blastp hits against human proteins ( $1 \times 10^{-6}$ ).

Repeats were annotated and filtered with RepeatMasker using a custom library generated with RepeatModeller. Long non-coding RNAs were identified by filtering all transcripts for protein-coding potential using CPAT<sup>43</sup> trained with zebrafish transcripts, and further discarding those that had a positive hit in a HMM search against the NR and PFAM databases (Extended Data Fig. 1g).

**Comparative genomics.** We used OMA<sup>44</sup> to reconstruct gene families and infer homology relationships based on well-established phylogenetic relationships between species<sup>45</sup>, and further merged families sharing Ensembl paralogues with ‘Euteleostomi’ or ‘Vertebrata’ ancestry. To define the set of high-confidence ohnologue families (Supplementary Data 2, dataset 9), we retained families with two to four copies in three out of five vertebrates (excluding teleosts) and subjected them to phylogenetic reconciliation.

To assess genome sequence conservation, reciprocal whole-genome alignments of *Branchiostoma floridae*, *Branchiostoma belcheri* and *B. lanceolatum* were performed using LASTZ and processed with phastCons<sup>46</sup> to produce conservation scores. The distribution of phastCons scores in APREs was determined using ‘dynamic’ ATAC-seq peaks that showed no temporal discontinuity in activity.

**Comparative transcriptomics.** To investigate the evolutionary conservation of chordate development at the molecular level, newly generated data from zebrafish, medaka and amphioxus, as well as available data from the SRA (frog and chicken), were compared (Supplementary Data 2, dataset 3 and Supplementary Table 3). Gene expression was estimated with Kallisto<sup>47</sup> using Ensembl transcriptome annotations (Supplementary Table 4), and summing up transcripts per million (TPMs) from all transcript isoforms to obtain one individual gene-expression estimate per sample. We used single-copy orthologues to pair genes and used the Jensen–Shannon distance metrics after quantile normalization of TPMs to score distance between pairs of transcriptomes:

$$JSD_s = \sqrt{\frac{1}{2} \sum_{g=0}^{n_{og}} p_g \times \log \left( \frac{p_g}{\frac{1}{2}(p_g + q_g)} \right) + \frac{1}{2} \sum_{g=0}^{n_{og}} q_g \times \log \left( \frac{q_g}{\frac{1}{2}(p_g + q_g)} \right)}$$

Statistical robustness towards gene sampling was assessed by calculating transcriptomic distances based on 100 bootstrap replicates and estimating the standard deviation over these replicates.

To obtain groups of genes with similar dynamics of expression during development, genes were clustered based on their cRPKM<sup>48</sup> using the Mfuzz package<sup>49</sup>. For this purpose, eight comparable stages were selected in amphioxus and zebrafish on the basis of conserved developmental landmarks such as fertilization,

gastrulation and organogenesis (Supplementary Table 5). The statistical significance of the orthologous gene overlap between pairs of clusters was assessed using upper-tail hypergeometric tests.

Modules of co-expressed genes across stages and adult tissues were inferred using WGCNA<sup>50</sup> with default parameters in amphioxus (17 samples) and zebrafish (27 samples) (Supplementary Table 6). The statistical significance of the orthologous gene overlap between pairs of clusters was assessed using upper-tail hypergeometric tests. The numbers of transcription-factor binding-site motifs detected in APREs in the basal regions of genes from any given cluster were standardized using z-scores.

To have a general assessment of the extent of conservation or divergence in gene expression among chordates at adult stages, we used neighbourhood analysis of conserved co-expression (NACC)<sup>25</sup>, a method developed to compare heterogeneous, non-matched sample sets across species. NACC relies on comparisons of average distances between pairs of orthologous (genes A and B), the 20 genes with the closest transcriptomic distance ( $\bar{A}$  and  $\bar{B}$ ) and their reciprocal orthologues in the other species ( $\bar{AB}$  and  $\bar{BA}$ ), and is calculated as follows:

$$NACC = \frac{1}{2} [(\bar{AB} - \bar{A}) + (\bar{BA} - \bar{B})]$$

NACC calculations were performed for each family that contained a single amphioxus member and up to eight members in zebrafish and were also performed with randomized orthology relationships as a control.

**Regulatory profiling.** ATAC-seq. For amphioxus, medaka and zebrafish, ATAC-seq was performed in two biological replicates by directly transferring embryos in the lysis buffer, following the original protocol<sup>51,52</sup>. ATAC-seq libraries were sequenced to produce an average of 66, 83 and 78 million reads for amphioxus, zebrafish and medaka, respectively. Reads were mapped with Bowtie2 and nucleosome-free pairs (insert < 120 bp) retained for peak-calling using MACS2<sup>53</sup>, and the irreducible discovery rate was used to assess replicability. Nucleosome positioning was calculated from aligned ATAC-seq data using NucleoATAC<sup>54</sup>.

**Chromatin immunoprecipitation with sequencing (ChIP-seq).** Embryos of undetermined gender were fixed in 2% formaldehyde and ChIP was performed as previously described for other species<sup>55</sup>. Chromatin was sonicated and incubated with the corresponding antibody (H3K4me3: ab8580, H3K27ac: ab4729 and H3K27me3: ab6002, from Abcam). An average of 30 million reads per library was generated. Reads were mapped with Bowtie2 and peaks called with MACS2<sup>53</sup>, assuming default parameters.

**4C-seq.** Embryos of undetermined gender were fixed in 2% formaldehyde and chromatin was digested with DpnII and Csp6. Specific primers targeted the TSSs of the studied genes and included Illumina adapters. An average 5 million reads were generated for each of the two biological replicates. After mapping, reads were normalized per digestion fragment cut and interactions were identified using peakC<sup>56</sup> with low-coverage regions excluded.

**MethylC-seq and RRBS.** Genomic DNA was extracted as previously described<sup>57</sup>, sonicated, purified and end-repaired. Bisulfite conversion was performed with the MethylCode Bisulfite Conversion Kit (Thermo Fisher Scientific). After Illumina library construction, an average of 73 million reads per sample were sequenced. RRBS libraries were prepared similarly to those for MethylC-seq, but with restriction digestion with MspI instead of sonication and PCR amplification. An average of 46 million reads per sample was generated. Reads were mapped to an in silico, bisulfite-converted *B. lanceolatum* reference genome<sup>7,58</sup>. Differentially methylated regions in the CpG context were identified as previously described<sup>7</sup>. Differential transcription-factor motif enrichment was obtained with DiffBind from Bioconductor.

**CAGE-seq.** Libraries were constructed using the non-amplifying non-tagging Illumina CAGE protocol<sup>59</sup>. Mouse CAGE-seq data were obtained from FANTOM5<sup>60</sup>. Reads were aligned using Bowtie. Nearby individual CAGE TSSs were combined using the distance-based clustering method in CAGEr<sup>61</sup> to produce tag clusters, which summarize expression at individual promoters. Tag clusters were clustered across samples to produce comparable promoter regions, referred to as ‘consensus clusters’. The consensus clusters were then grouped by expression patterns using a self-organizing map<sup>62</sup>. We investigated the relative presence and enrichment of the following features: TATA box, YY1 motif, GC and AT content, SS and WW dinucleotides, first exons and nucleosome positioning signal. Heat maps were plotted for visualization by scanning either for exact dinucleotide matches or for position weight matrix matches at 80% of the maximum score. Position weight matrices for TATA and YY1 were taken from the JASPAR vertebrate collection.

**Cis-regulatory comparisons.** Depending on the analysis, an APRE was associated with a specific gene if it was located within: (i) the ‘basal’ region of the gene (–5 kb to +1 kb of the TSS; for comparisons of enriched motif composition) or (ii) the GREAT region of the gene (up to ±1 Mb of the TSS unless another basal region was found; for comparing the number of APREs per gene)<sup>26</sup>. Stratification of gene

sets by GREAT or intergenic-region size between amphioxus and zebrafish was done using the function stratify from the matt suite<sup>63</sup>, with a range of  $\pm 500$  bp.

The DNA-binding specificity of each transcription factor was predicted on the basis of the binding domain similarity to other transcription-factor family members, as previously performed<sup>64</sup>. Transcription-factor motifs from CIS-BP version 1.02<sup>64</sup> were downloaded and clustered using GimmeMotifs<sup>65</sup> ( $P \leq 0.0001$ ). Two hundred and forty-two clusters of motifs were assigned to one or more orthologous groups in both amphioxus and zebrafish and used for all analyses (Supplementary Data 2, dataset 10). These motifs were detected in APREs using the tools gimme threshold and gimme scan from GimmeMotifs<sup>65</sup>.

**Effect of WGDs on gene expression.** Gene expression was binarized (1 if the normalized cRPKM  $> 5$ , and 0 otherwise) across nine comparable samples in amphioxus and three vertebrate species (mouse, frog and zebrafish) (Supplementary Table 7). Then, for each amphioxus gene and vertebrate orthologue, the expression bias was measured by subtracting the number of positive-expression domains in amphioxus from that of vertebrates (Fig. 6a). The amphioxus gene-expression pattern was also compared to the union of the orthologues, as well as the pattern after binarizing the expression for the sum of cRPKM values of all family members. The analysis was restricted to families with a single member in amphioxus.

Next, we selected those orthologue families for which the ancestral expression included the nine studied domains, as inferred from having expression in the single amphioxus orthologue and in the union of the family. For each gene family, we then defined (Fig. 6c): (i) redundancy (all vertebrate paralogues were expressed in all domains), (ii) subfunctionalization (none of the vertebrate members had expression across all domains<sup>27</sup>), and (iii) specialization (one or more vertebrate orthologues were expressed in all domains, but at least one orthologue was not). Members of the later type were subdivided into 'strong' and 'mild' specialization if they retained  $\leq 2$  or more expression domains. We examined the transcript sequence similarity as well as the dN/dS between human and mouse (retrieved from Biomart), and the number of APREs associated with genes from different categories. Finally, we computed the  $\tau$  tissue-specificity index as previously described<sup>28</sup>, to assess more broadly the tissue specificity of orthologues.

**Transgenic assays in zebrafish and amphioxus.** Enhancer reporter assays in zebrafish embryos were performed as previously described<sup>66</sup>. Selected peaks were first amplified, cloned into a PCR8/GW/TOPO vector and transferred into a detection vector (including a *gata2* minimal promoter, a GFP reporter gene and a strong midbrain enhancer (z48) as an internal control)<sup>67</sup>. Transgenic embryos were generated using the Tol2 transposon and transposase method<sup>68</sup>. Three or more independent stable transgenic lines were generated for each construct as reported in Supplementary Table 8. For amphioxus reporter assays, selected peaks were amplified and transferred into a detection vector (including the *Branchiostoma* minimal actin promoter, a GFP reporter gene and piggyBac terminal repeats). Transgenic embryos were generated by the piggyBac transposase method.

**In situ hybridization.** Gene fragments that were synthetically designed or amplified by PCR from cDNA were sub-cloned into pBluescript II SK and used as templates for probe synthesis using the DIG labelling kit (Roche) and T3 RNA polymerase. Embryos at different developmental stages were fixed in PFA 4% dissolved in MOPS-EGTA buffer and in situ hybridization carried out as previously described<sup>69</sup>, using BCIP/NBT as a chromogenic substrate.

**Reporting summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

**Code availability.** Custom code is available at <https://gitlab.com/groups/FunctionalAmphioxus>.

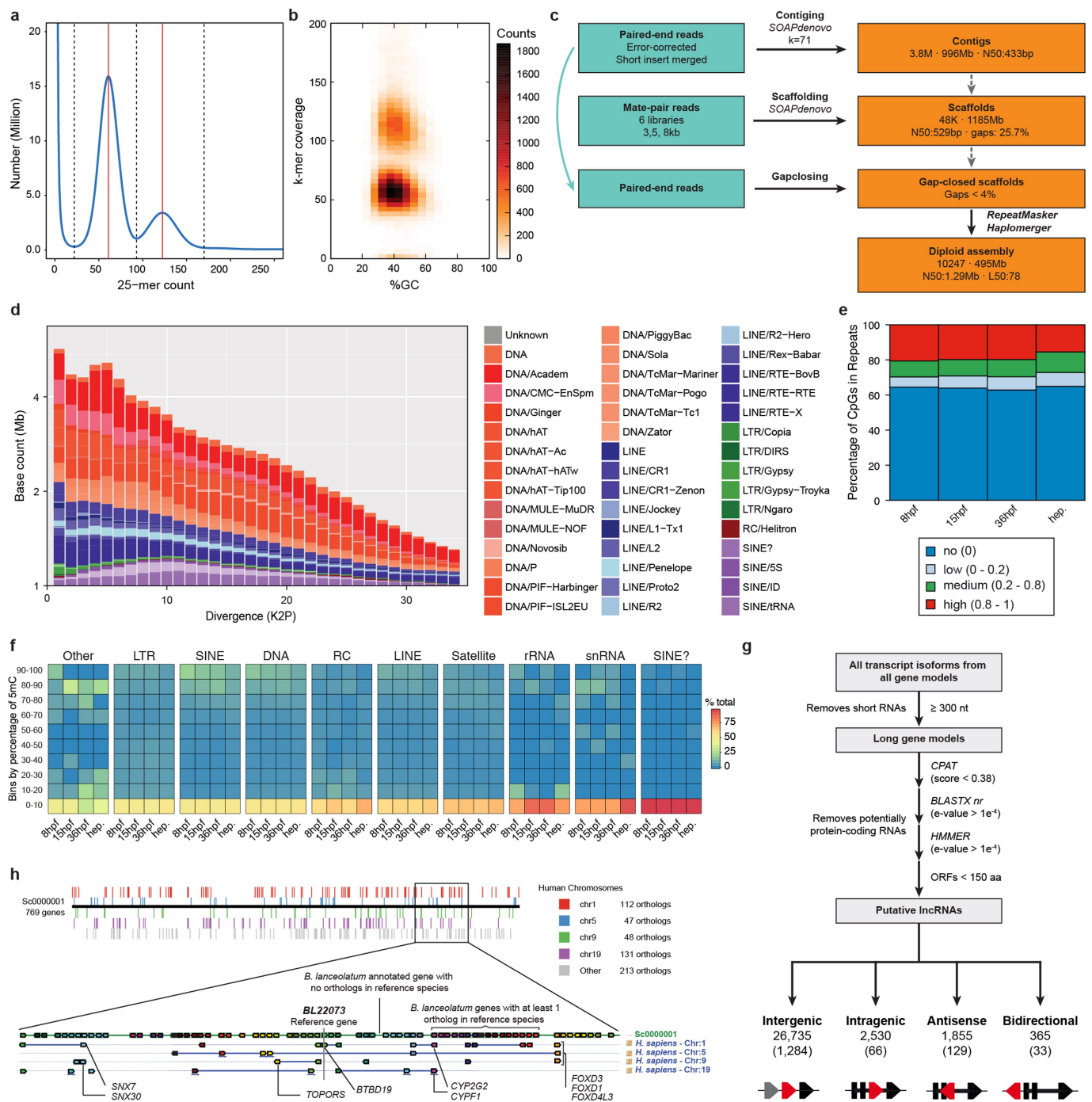
## Data availability

Next-generation sequencing data have been deposited in Gene Expression Omnibus (GEO) under the following accession numbers: GSE106372 (ChIP-seq), GSE106428 (ATAC-seq), GSE106429 (CAGE-seq), GSE106430 (RNA-seq), GSE102144 (MethylC-seq and RRBS) and GSE115945 (4C-seq). Raw genome sequencing data and the genome assembly have been submitted to European Nucleotide Archive (ENA) under the accession number PRJEB13665. UCSC hub and annotation files are available at <http://amphiencode.github.io/>.

30. Fuentes, M. et al. Preliminary observations on the spawning conditions of the European amphioxus (*Branchiostoma lanceolatum*) in captivity. *J. Exp. Zool. B Mol. Dev. Evol.* **302B**, 384–391 (2004).
31. Fuentes, M. et al. Insights into spawning behavior and development of the European amphioxus (*Branchiostoma lanceolatum*). *J. Exp. Zool. B Mol. Dev. Evol.* **308B**, 484–493 (2007).
32. Hirakow, R. & Kajita, N. Electron microscopic study of the development of amphioxus, *Branchiostoma belcheri tsingtauense*: the gastrula. *J. Morphol.* **207**, 37–52 (1991).
33. Hirakow, R. & Kajita, N. Electron microscopic study of the development of amphioxus, *Branchiostoma belcheri tsingtauense*: the neurula and larva. *Kaibogaku Zasshi* **69**, 1–13 (1994).

34. Luo, R. et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* **1**, 18 (2012).
35. Huang, S. et al. HaploMerger: reconstructing allelic relationships for polymorphic diploid genome assemblies. *Genome Res.* **22**, 1581–1588 (2012).
36. Grabherr, M. G. et al. Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
37. Haas, B. J. et al. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666 (2003).
38. Keller, O., Kollmar, M., Stanke, M. & Waack, S. A novel hybrid gene prediction method employing protein multiple sequence alignments. *Bioinformatics* **27**, 757–763 (2011).
39. Haas, B. J. et al. Automated eukaryotic gene structure annotation using EvidenceModeler and the program to assemble spliced alignments. *Genome Biol.* **9**, R7 (2008).
40. Kim, D. et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36 (2013).
41. Trapnell, C. et al. Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).
42. Haas, B. J. et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protocols* **8**, 1494–1512 (2013).
43. Wang, L. et al. CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Res.* **41**, e74 (2013).
44. Roth, A. C., Gonnet, G. H. & Dessimoz, C. Algorithm of OMA for large-scale orthology inference. *BMC Bioinformatics* **9**, 518 (2008).
45. Altenhoff, A. M., Gil, M., Gonnet, G. H. & Dessimoz, C. Inferring hierarchical orthologous groups from orthologous gene pairs. *PLoS ONE* **8**, e53786 (2013).
46. Siepel, A. et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).
47. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).
48. Labbé, R. M. et al. A comparative transcriptomic analysis reveals conserved features of stem cell pluripotency in planarians and mammals. *Stem Cells* **30**, 1734–1745 (2012).
49. Kumar, L. & Futschik, M. E. Mfuzz: a software package for soft clustering of microarray data. *Bioinformatics* **2**, 5–7 (2007).
50. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).
51. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).
52. Fernández-Miñán, A., Bessa, J., Tena, J. J. & Gómez-Skarmeta, J. L. Assay for transposase-accessible chromatin and circularized chromosome conformation capture, two methods to explore the regulatory landscapes of genes in zebrafish. *Methods Cell Biol.* **135**, 413–430 (2016).
53. Zhang, Y. et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
54. Schep, A. N. et al. Structured nucleosome fingerprints enable high-resolution mapping of chromatin architecture within regulatory regions. *Genome Res.* **25**, 1757–1770 (2015).
55. Bogdanović, O., Fernández-Miñán, A., Tena, J. J., de la Calle-Mustienes, E. & Gómez-Skarmeta, J. L. The developmental epigenomics toolbox: ChIP-seq and MethylCap-seq profiling of early zebrafish embryos. *Methods* **62**, 207–215 (2013).
56. Geeven, G., Teunissen, H., de Laat, W. & de Wit, E. peakC: a flexible, non-parametric peak calling package for 4C and Capture-C data. *Nucleic Acids Res.* **46**, e91 (2018).
57. Bogdanović, O. & Veenstra, G. J. Affinity-based enrichment strategies to assay methyl-CpG binding activity and DNA methylation in early *Xenopus* embryos. *BMC Res. Notes* **4**, 300 (2011).
58. Lister, R. et al. Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. *Nature* **471**, 68–73 (2011).
59. Murata, M. et al. Detecting expressed genes using CAGE. *Methods Mol. Biol.* **1164**, 67–85 (2014).
60. The FANTOM Consortium and the RIKEN PMI and CLST (DGT). A promoter-level mammalian expression atlas. *Nature* **507**, 462–470 (2014).
61. Haberle, V., Forrest, A. R., Hayashizaki, Y., Carninci, P. & Lenhard, B. CAGER: precise TSS data retrieval and high-resolution promoterome mining for integrative analyses. *Nucleic Acids Res.* **43**, e51 (2015).
62. Wehrens, R. & Buydens, L. M. C. Self- and super-organising maps in R: the kohonen package. *J. Stat. Softw.* **21**, 1–19 (2007).
63. Gohr, A. & Irimia, M. Matt: Unix tools for alternative splicing analysis. *Bioinformatics* (2018).
64. Weirauch, M. T. et al. Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**, 1431–1443 (2014).
65. van Heeringen, S. J. & Veenstra, G. J. GimmeMotifs: a de novo motif prediction pipeline for ChIP-sequencing experiments. *Bioinformatics* **27**, 270–271 (2011).
66. Bessa, J. et al. Zebrafish enhancer detection (ZED) vector: a new tool to facilitate transgenesis and the functional analysis of cis-regulatory regions in zebrafish. *Dev. Dyn.* **238**, 2409–2417 (2009).
67. Gehrke, A. R. et al. Deep conservation of wrist and digit enhancers in fish. *Proc. Natl Acad. Sci. USA* **112**, 803–808 (2015).

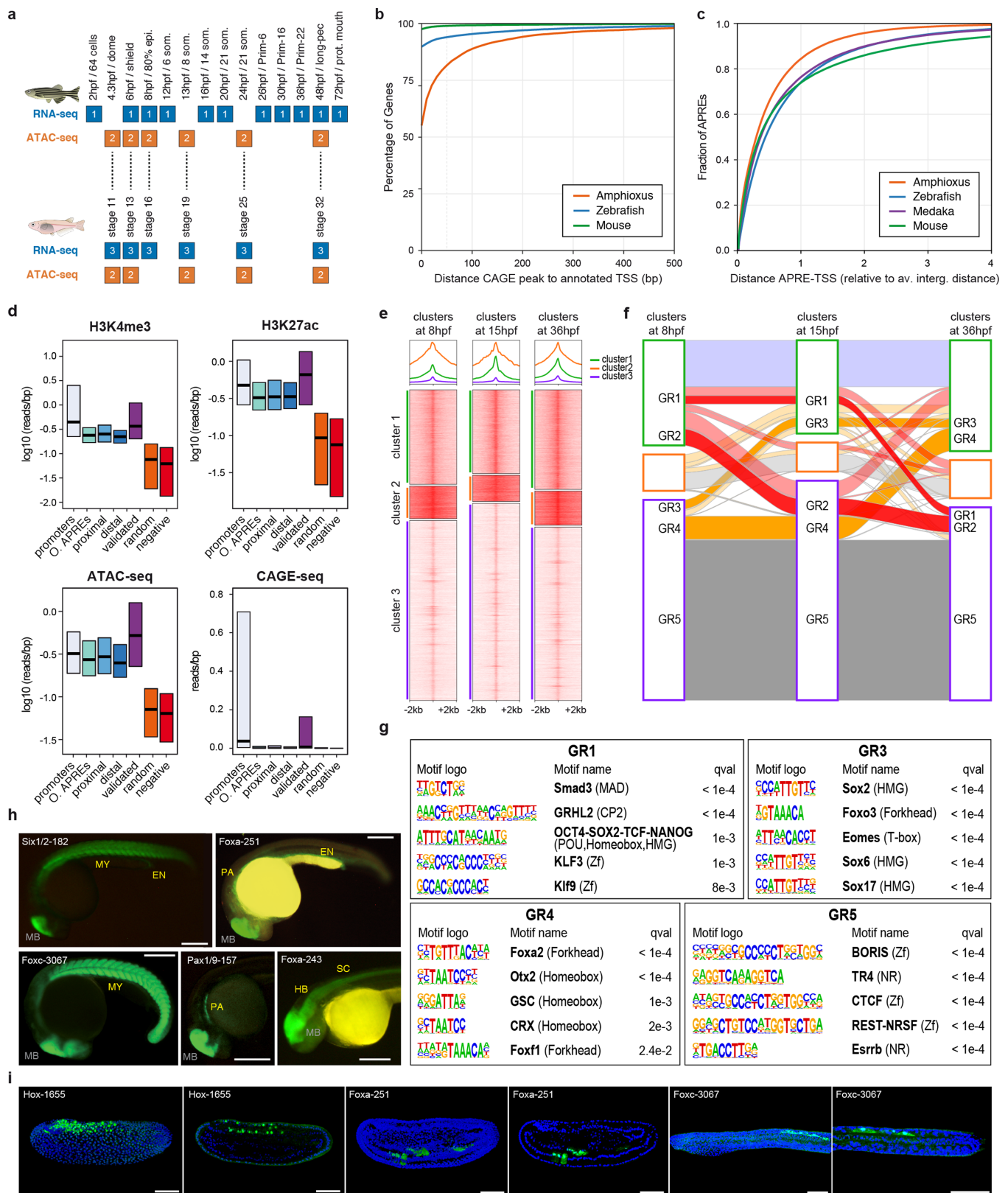
68. Kawakami, K. Transgenesis and gene trap methods in zebrafish by using the *Tol2* transposable element. *Methods Cell Biol.* **77**, 201–222 (2004).
69. Somorjai, I., Bertrand, S., Camasses, A., Haguénauer, A. & Escriva, H. Evidence for stasis and not genetic piracy in developmental expression patterns of *Branchiostoma lanceolatum* and *Branchiostoma floridae*, two amphioxus species that have evolved independently over the course of 200 Myr. *Dev. Genes Evol.* **218**, 703–713 (2008).
70. Tena, J. J. et al. Comparative epigenomics in distantly related teleost species identifies conserved *cis*-regulatory nodes active during the vertebrate phylotypic period. *Genome Res.* **24**, 1075–1085 (2014).
71. Acemel, R. D. et al. A single three-dimensional chromatin compartment in amphioxus indicates a stepwise evolution of vertebrate Hox bimodal regulation. *Nat. Genet.* **48**, 336–341 (2016).



**Extended Data Fig. 1 | Summary of genomic assembly and repeat annotation.** **a**, Spectrum of 25-mers in Illumina sequencing data that shows the bimodal distribution that is characteristic of highly polymorphic species. **b**, Heat map showing *k*-mer decomposition (y axis) across GC content (x axis). Both peaks show comparable GC content, which is consistent with them representing haploid versus diploid *k*-mers. **c**, Flow chart of the steps followed to obtain the *B. lanceolatum* assembly. **d**, Repeat landscape and its evolutionary history, shown by the proportion of repetitive elements with a given divergence (K2P) for their consensus in the repeat library (repeatScout). **e**, Percentage of methylated CpG dinucleotides within repetitive elements, at three developmental stages and in the adult hepatic diverticulum. **f**, Distribution of average levels of 5mC of different repeat families. Colour key indicates the percentage of repeats in each family with corresponding levels of average methylation. **g**, Computational pipeline to identify long non-coding RNAs (lncRNAs). Categories: antisense, lncRNA overlaps with a protein-coding gene in the reverse strand; intragenic, lncRNA overlaps with a protein-coding gene in the same strand; bidirectional, within 1 kbp of a TSS of a protein-coding gene in the antisense strand, probably a product of a

bidirectional promoter; intergenic, lncRNA does not overlap with any protein-coding gene. The total number in each category is indicated, with the number of those that are multi-exonic in parentheses. **h**, Quadruple conserved synteny between amphioxus and human. Top, amphioxus scaffold Sc00000001 aligned against the four human chromosomes with which it shares the highest number of orthologues (chr1, chr5, chr9 and chr19). In this scaffold, 277 out of 551 genes have clear orthologues in human, and 203 of these have orthologues on at least one of the four mentioned chromosomes. The black horizontal line represents the amphioxus scaffold, and each vertical coloured box an orthologous gene on the corresponding human chromosome. Bottom, modified view from Genomicus that is centred on the *BL22073* gene and spans Sc00000001: 7,736,434–8,850,041. On the top line, each amphioxus gene with at least one orthologue in the nine reference species is represented with an oriented coloured box. Human genes located in the four orthologous chromosomes are aligned underneath, in boxes of colours that correspond to those of their amphioxus pro-orthologues. The Genomicus server dedicated to amphioxus can be accessed at <http://genomicus.biologie.ens.fr/genomicus-amphioxus>.

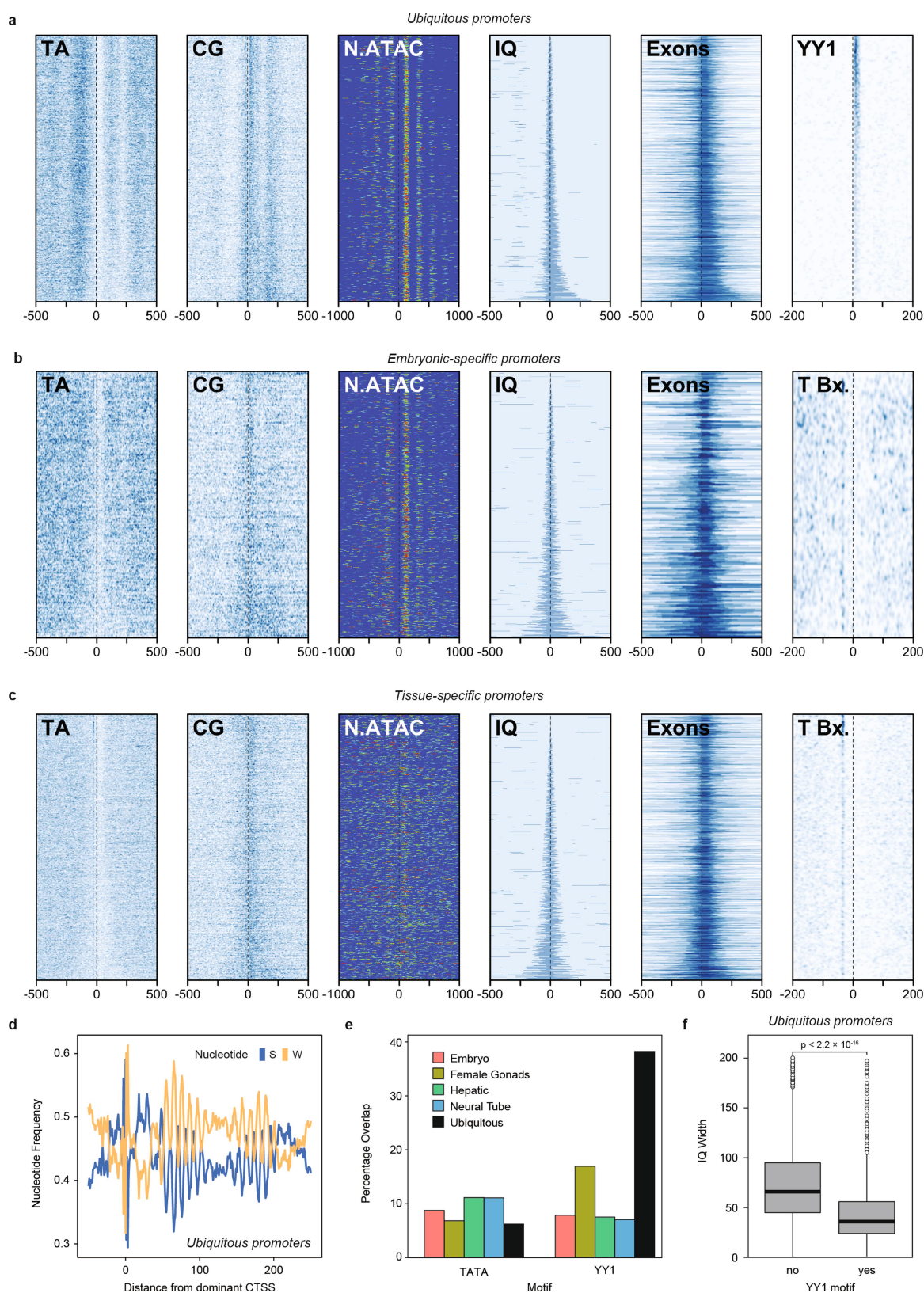




Extended Data Fig. 2 | See next page for caption.

**Extended Data Fig. 2 | Dynamics of chromatin marks on APREs and reporter assays.** **a**, Summary of the zebrafish and medaka RNA-seq and ATAC-seq datasets generated for this study. Dashed lines indicate equivalent developmental stages in the two species, based on a previous study<sup>70</sup>. The number of biological replicates is indicated for each experiment. Zebrafish 24-hpf ATAC-seq data are from a previous study<sup>67</sup>. **b**, Cumulative distribution of the distance between CAGE-seq peaks and the closest annotated TSSs for genes with expression cRPKM > 5 in any of the samples covered by CAGE-seq (see Fig. 1a). Only CAGE-seq peaks within 1 kbp of an annotated TSS were tested (amphioxus: 10,435 peaks; zebrafish, 23,326 peaks; and mouse, 23,443 peaks). **c**, Cumulative distribution of distances between each APRE and the closest annotated TSS normalized by the average intergenic distance of the species (amphioxus, 83,471; zebrafish, 252,774; medaka, 174,139; and mouse, 216,857 APREs, as per Fig. 1c). **d**, Signal distribution of different marks within functional-genomic regions in amphioxus. log<sub>10</sub> of read counts of H3K4me3, H3K27ac and ATAC-seq, and raw read counts of CAGE-seq in promoters of homology-supported, protein-coding genes ( $n = 26,501$ ), other APREs ('O. APREs', all APREs that do not overlap a TSS from any gene model;  $n = 48,341$ ), proximal APREs ( $n = 24,622$ ), distal APREs ( $n = 11,881$ ), previously validated enhancers ( $n = 43$ ; Supplementary Table 9), random regions ( $n = 88,413$ ) and negative regions (excluding ATAC-seq peaks,  $n = 88,413$ ). For region designation, see Fig. 1c. For clarity, whiskers and outliers are not displayed. **e**, *k*-means clustering of APREs based on H3K27ac signal in three developmental stages. Cluster 1 and 3 APREs were considered as active and inactive, respectively. Average H3K27ac profiles are represented in the top panels. The number of APREs per cluster and stage are provided in Supplementary Data 2, dataset 8. **f**, Alluvial plot that shows the dynamics of each APRE among

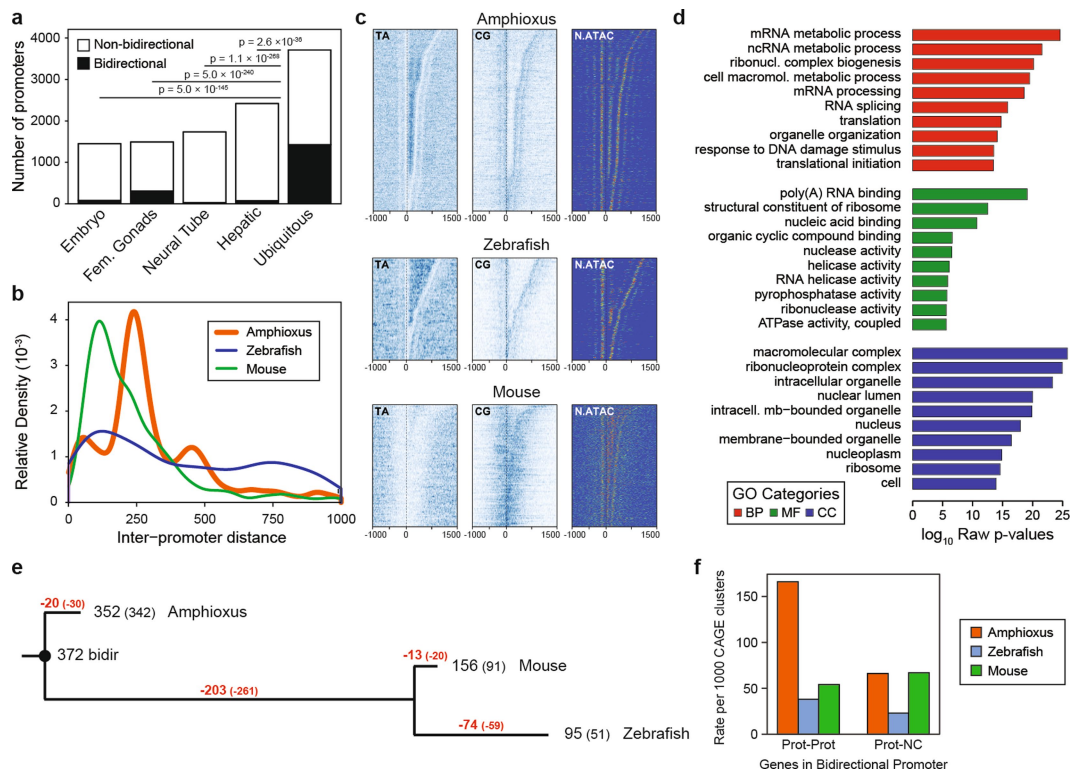
the clusters described in **e**. APREs that remained active (cluster 1 in all stages) along the three developmental stages are represented in blue, constitutively inactive APREs (cluster 3 in all stages) in dark grey and dynamic APREs in red or orange (if inactivated or activated, respectively, during development). Five groups of APREs of special interest are highlighted with stronger colours and named GR1–GR5. **g**, Representative enriched DNA motifs found in each of the groups described in **f**. GR1 APREs were enriched in early motifs (for example, Smad3 and Oct4, Sox2 and Nanog); GR3 APREs in motifs of transcription factors involved in the generation of the three germ layers (for example, Foxo3, Sox6 and Sox17); GR4 APREs in tissue-specific transcription factors (for example, Foxa2, Otx2 and Crx); and GR5 APREs in CTCF and CTCF-like (BORIS) motifs. *q* values as provided by Homer. **h**, Lateral views of embryos from stable transgenic zebrafish lines at 24 hpf (except for Foxa-243, at 48 hpf) showing GFP expression driven by the amphioxus APREs listed in Supplementary Table 8 and highlighted in Supplementary Fig. 1. The number of independent founders with the same expression were as follows: Six1/2-182 (5/5), Foxa-243 (3/3), Foxa-251 (4/4), FoxC-3067 (6/6) and Pax1/9-157 (3/3). Midbrain expression corresponds to the positive-control enhancer included in the reporter constructs. EN, endoderm; HB, hindbrain; MY, myotomes; PA, pharyngeal arch; SC, spinal cord. Scale bar, 250  $\mu$ m. **i**, Lateral views of transient transgenic amphioxus embryos, showing GFP expression driven by the APREs highlighted in Supplementary Fig. 1a, b (Foxa-251 ( $n = 46$  out of 52) and Foxc-3067 ( $n = 27$  out of 35), respectively) and in a previous study<sup>71</sup> (Hox-1655,  $n = 72$  out of 80). For each element, left panels correspond to 3D rendering from sub-stacks and right panels to *z*-stack sagittal sections. Scale bar, 50  $\mu$ m. Anterior is to the left and dorsal to the top.



**Extended Data Fig. 3 | Features of amphioxus promoters derived from CAGE-seq.** **a–c**, Heat maps showing AT and CG signal, nucleosome positioning (derived from the NucleoATAC signal), promoter width (interquantile (IQ) range), first exon length and YY1 (a) or TATA box (b, c) motifs around ubiquitous (a,  $n = 3,710$ ), embryonic-specific (b,  $n = 1,451$ ) and tissue-specific (c,  $n = 4,154$ ) promoters, sorted by promoter width. Position 0 corresponds to the main TSS. **d**, Ubiquitous promoters show strong evidence for a nucleosome positioned downstream

of the CAGE TSS, as judged from the 12-bp periodicity of W and S nucleotide density. **e**, Per cent of promoters of each category that have associated TATA box or YY1 motifs. Number of promoters: embryo, 1,451; female gonads, 1,494; hepatic, 2,420; neural tube, 1,734; and ubiquitous, 3,710. **f**, IQ width distribution of ubiquitous promoters ( $n = 3,710$ ) with and without an associated YY1 motif.  $P$  value corresponds to two-sided Wilcoxon sum-rank tests.

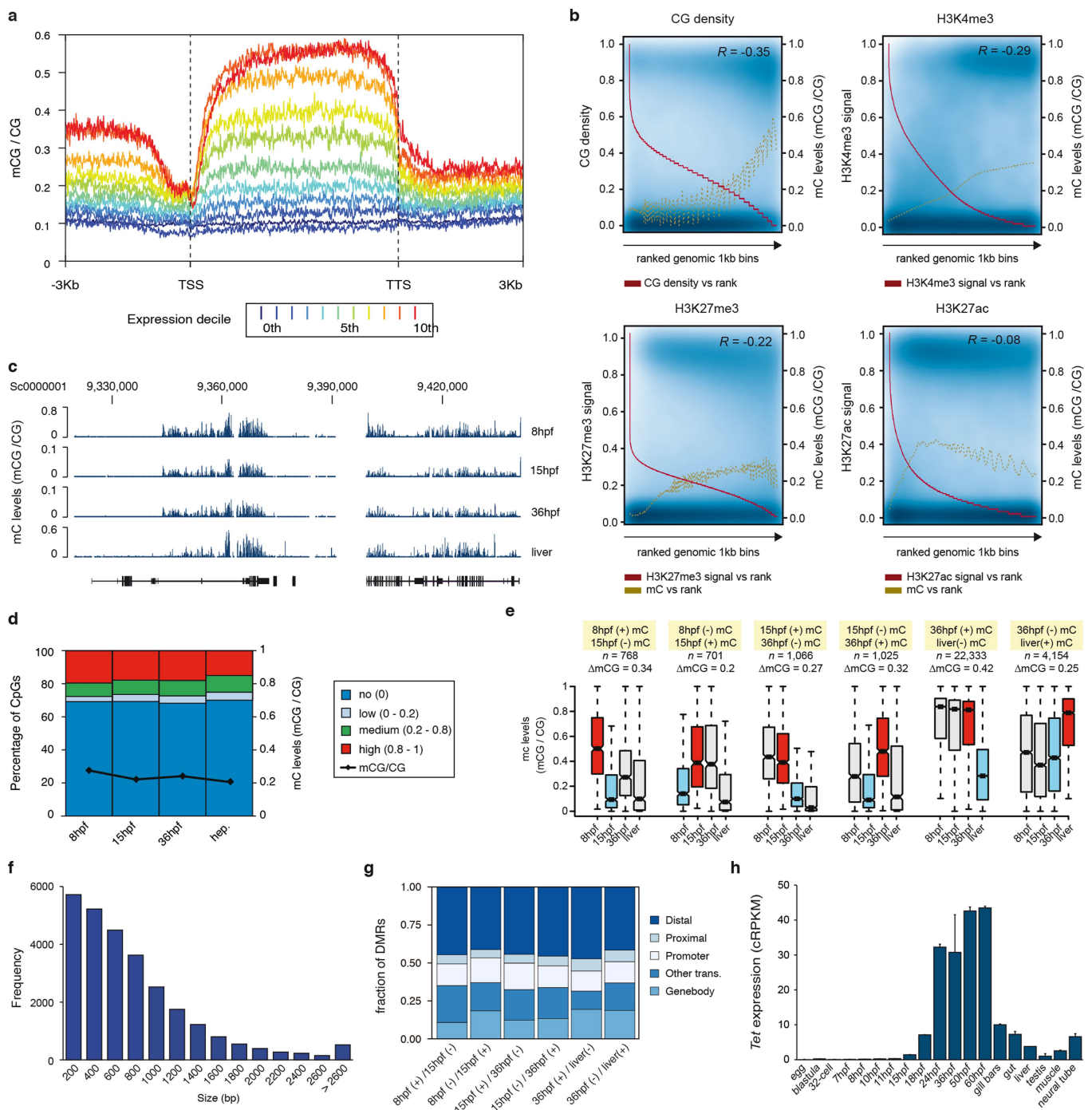




**Extended Data Fig. 4 | Characteristics and evolution of bidirectional promoters.** **a**, Number of bidirectional and non-bidirectional promoters identified for each regulatory category.  $P$  values correspond to two-sided Fisher's exact tests against ubiquitous promoters. **b**, Distribution of distance between bidirectional promoters in each species (amphioxus, 1,975; zebrafish, 549; and mouse, 876 pairs of promoters). The distance between amphioxus peaks closely corresponds to integral nucleosome spacing. **c**, Heat maps of TA, CG and nucleosome occupancy (derived from the NucleoATAC signal) around bidirectional promoter pairs in amphioxus ( $n = 1,975$ ), mouse ( $n = 876$ ) and zebrafish ( $n = 549$ ), arranged by the distance between the two CAGE TSSs. In amphioxus, both TA and NucleoATAC signals indicate regions in which 0, 1 or 2 nucleosomes separate promoters. **d**, Enriched GO terms for genes associated with bidirectional promoters in amphioxus. Uncorrected  $P$  values correspond to two-sided Fisher's exact tests as provided by topGO.

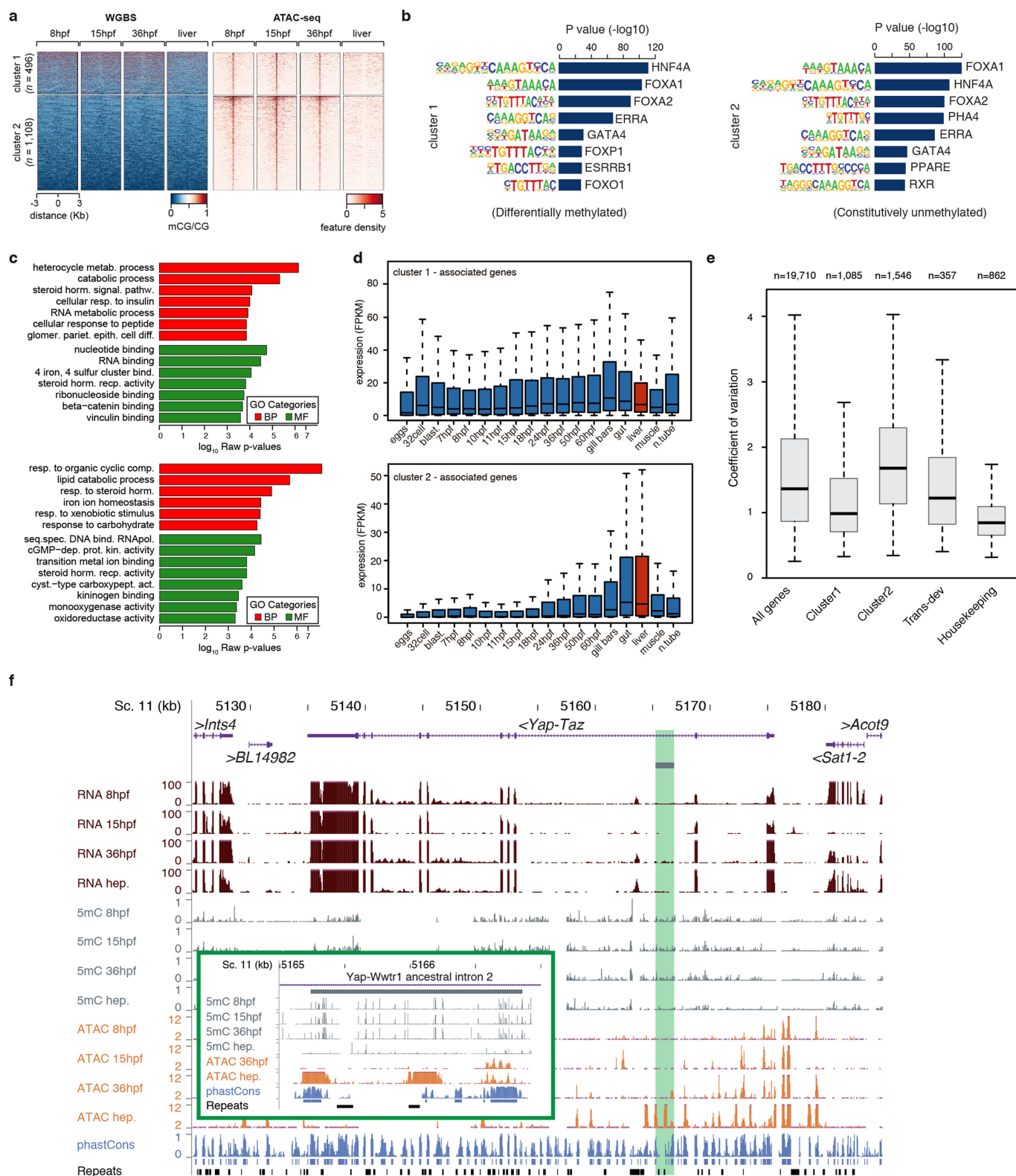
**e**, Inferred evolutionary dynamics of 372 putatively ancestral bidirectional promoters among chordate groups. Red, number of inferred losses and disentanglements; black, number of detected bidirectional promoters by CAGE-seq (in brackets) or microsynteny (neighbouring genes in a 5' to 5' orientation) for each species. In parentheses, number of lost and disentangled (red) or retained (black) bidirectional promoters when considering only the cases supported by CAGE-seq. **f**, In vertebrates, disentanglement was not accompanied by a general increase in the fraction of bidirectional promoters with antisense non-coding transcription, as shown by the relative number of CAGE clusters identified as bidirectional promoters that are composed of two protein-coding genes ('Prot-Prot') or of one protein-coding and one non-coding or non-annotated locus ('Prot-NC'). The total number of uniquely annotated, protein-coding-associated CAGE promoters was amphioxus, 11,789; mouse, 13,654; and zebrafish, 14,014.





**Extended Data Fig. 5 | 5mC dynamics in amphioxus.** **a**, 5mC levels across gene bodies ( $n = 20,569$ ) from different expression deciles (0<sup>th</sup>, not expressed; 10<sup>th</sup>, highest expression). TTS, transcription termination site. **b**, Scatter plots of levels of 5mC and CpG density, H3K4me3, H3K27me3 and H3K27ac in 1-kbp genomic bins sorted on the basis of feature rank. The red line represents a smoothing spline of 5mC signal versus feature rank number. Pearson correlation coefficients ( $R$ ) are displayed in the top right corner of each panel. **c**, UCSC browser excerpt of 5mC patterns for selected regions. **d**, Percentage of methylated CpG dinucleotides in 8-hpf ( $n = 19,657,388$ ), 15-hpf ( $n = 21,247,615$ ), 36-hpf ( $n = 21,702,000$ ) and hepatic (adult,  $n = 19,240,245$ ) amphioxus samples. Black line indicates the fraction between methylated and non-methylated CpGs at each stage. **e**, Box

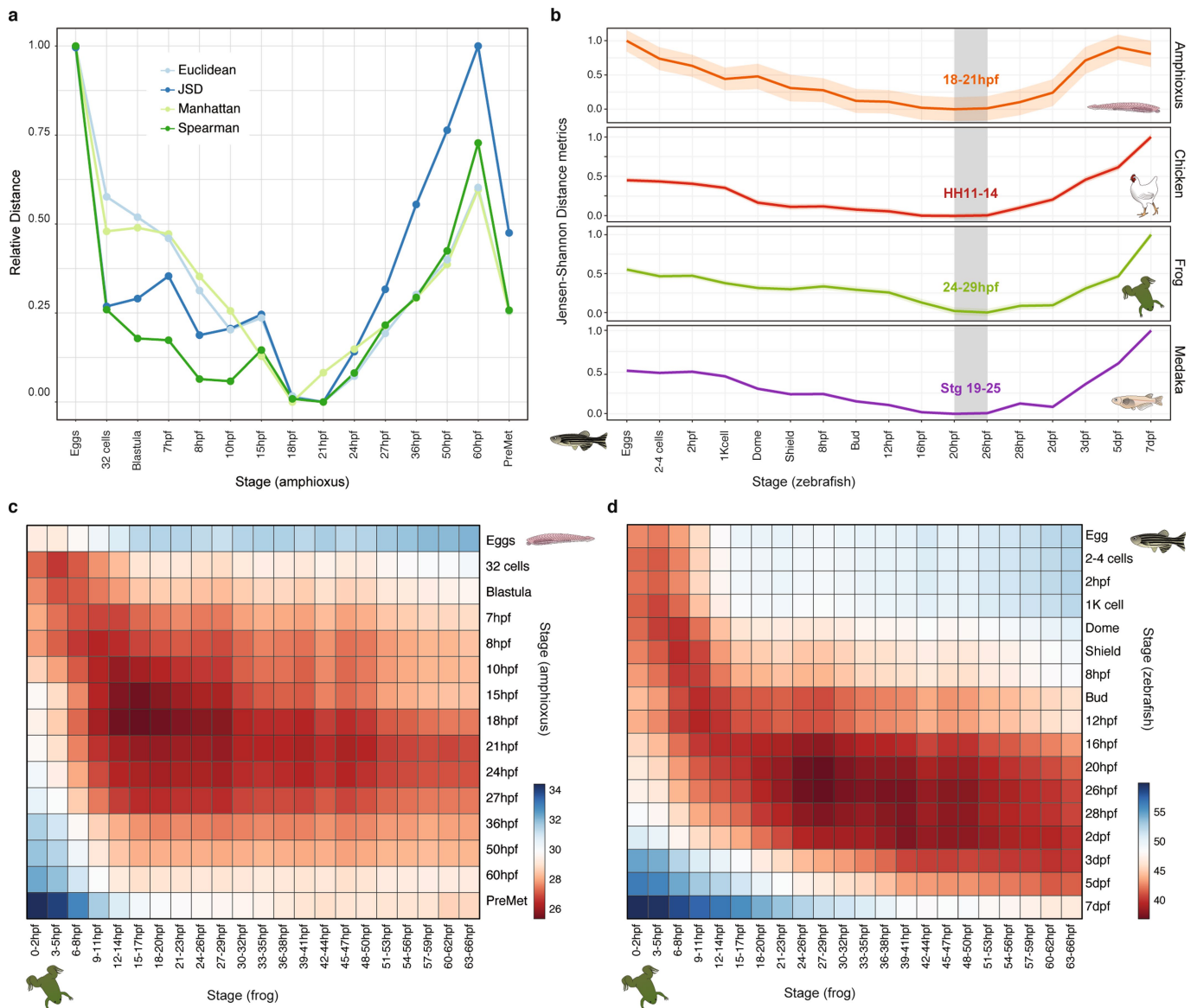
plots of average 5mC levels in different types of differentially methylated regions (DMRs) at each stage.  $\Delta$ mCG denotes the change in the fraction of methylated CpGs between the two stages used for identification of DMRs (red (hyper) and blue (hypo) boxes). The number of DMRs were as follows: 8 hpf(+)-15 hpf(–), 768; 8 hpf(–)-15 hpf(+), 701; 15 hpf(+)-36 hpf(–), 1,066; 15 hpf(–)-36 hpf(+), 1,025; 36 hpf(+)-liver(–), 22,333; and 36 hpf(–)-liver(+), 4,154. The coordinates for all DMRs are provided in Supplementary Data 2, dataset 11. **f**, Distribution of DMR sizes (in bp). **g**, Genomic distribution of DMRs identified for each sample. 'Other trans.' DMRs that overlap with gene models that were not defined as being supported by orthology. **h**, Expression (cRPKM) of the amphioxus *Tet* orthologue in embryos and adult tissues. Error bars represent standard error of the mean (the number of replicates for each RNA-seq dataset is provided in Fig. 1a).



Extended Data Fig. 6 | See next page for caption.

**Extended Data Fig. 6 | Developmental 5mC dynamics at APREs in amphioxus.** **a**, *k*-means clustering ( $n = 2$ ) of 5mC signal over embryo-specific open-chromatin regions (that is, APREs), assessed by ATAC-seq (Supplementary Table 10). **b**, The most significantly enriched transcription-factor binding-site motifs in APREs that display different developmental 5mC patterns in Fig. 2b. Uncorrected *P* values as provided by MEME. All plotted motifs had Benjamini-corrected *q* values of 0. **c**, GO enrichment for genes associated with cluster 1 (top) or cluster 2 (bottom) APREs from Fig. 2b. Uncorrected *P* values correspond to two-sided Fisher's exact tests as calculated by topGO. **d**, Distribution of expression values (cRPKMs) across all samples for genes associated with cluster 1 (top,  $n = 1,114$ ) or cluster 2 (bottom,  $n = 1,594$ ) APREs from

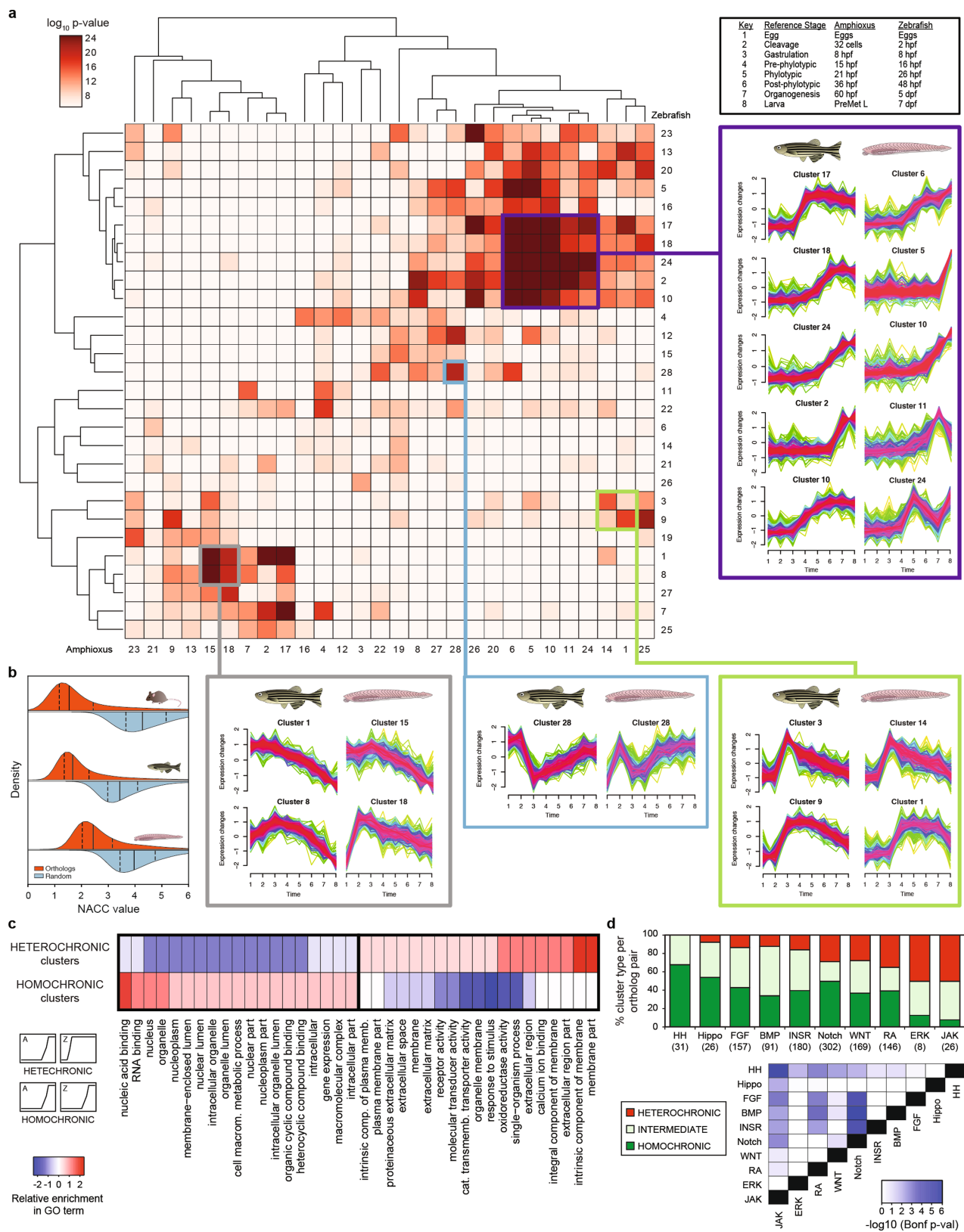
Fig. 2b. **e**, Distribution of the coefficients of variation for genes associated with cluster 1 or cluster 2 APREs from Fig. 2b, as well as all ( $n = 19,710$ ), trans-dev ( $n = 357$ ) and house-keeping ( $n = 862$ ) amphioxus genes. **f**, Example of a potentially conserved (zebrafish to amphioxus) DMR associated with *yap1*, a major transcription factor of the Hippo pathway. The inset corresponds to the region highlighted in green. The two ohnologous genomic regions in zebrafish are shown in Supplementary Fig. 2. Additional cases included genes that contained APREs that are likely to regulate neighbouring liver-specific genes ('bystander' genes) (Supplementary Table 11). The number of replicates for each experiment displayed in each track is provided in Fig. 1a.



**Extended Data Fig. 7 | Periods of maximal transcriptomic similarity across chordate development.** **a**, Stages of minimal transcriptomic distance obtained in the comparison between amphioxus and zebrafish for four alternative distance methods (Euclidean, Manhattan and Jensen–Shannon distances, and Spearman correlation). Values are normalized to minimal (0) and maximal (1) for each metric. **b**, Stages of minimal transcriptomic divergence shown as the smallest Jensen–Shannon distance between zebrafish stages and four chordate species. The shaded area surrounding the line that connects the stages is the standard deviation, derived from 100 bootstrap replicates of the orthologous gene set.

The grey box outlines the ‘phylogenetic’ period of minimal divergence; the corresponding periods are indicated for each species as the range provided by the two closest stages. **c**, **d**, Heat maps of pairwise transcriptomic distances (Jensen–Shannon distance metric) between pairs of chordate species, amphioxus and frog (**c**), and zebrafish and frog (**d**). In both heat maps, the smallest distance (red) indicates maximal similarity of the transcriptome. The periods of minimal divergence of the transcriptome are earlier for the amphioxus–frog comparison than for the zebrafish–frog comparison.

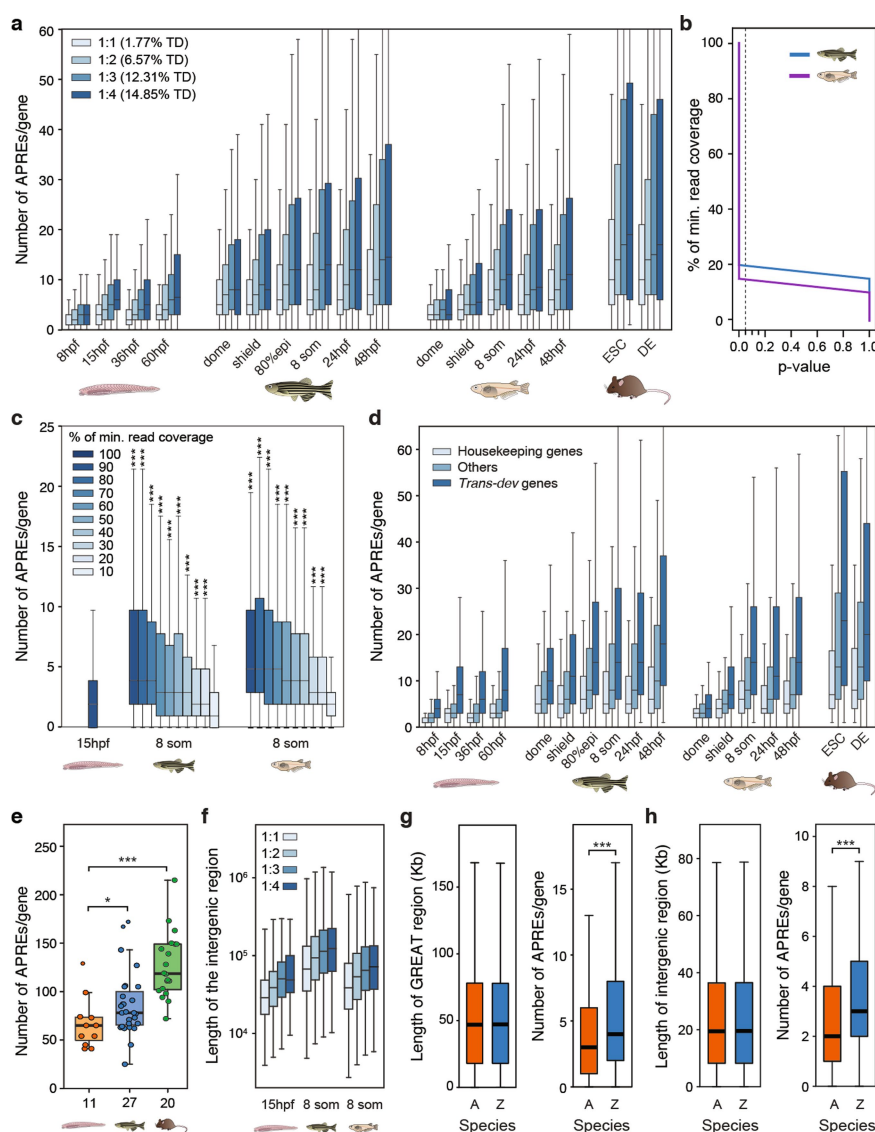




Extended Data Fig. 8 | See next page for caption.

**Extended Data Fig. 8 | Comparison of temporal gene expression profiles in amphioxus and zebrafish.** **a**, Heat map showing the significance of orthologous gene overlap between Mfuzz clusters across eight matched developmental stages in amphioxus and zebrafish as derived from an upper-tail hypergeometric test. Some clusters with highly significant overlap are highlighted, and their corresponding temporal expression profiles are shown. The profiles of all clusters for the two species are included in Supplementary Figs. 3, 4. Exact  $P$  values and sample sizes are provided in Supplementary Data 2, dataset 8. **b**, Distributions of NACC values for orthologous genes (in red) or random orthology assignments (blue) for each species against human. Lower NACC values imply higher conservation of relative expression. Solid lines show the median, and the dashed lines mark the interquartile range. The number of orthologue pairs were as follows: mouse, 15,109; zebrafish, 16,480;

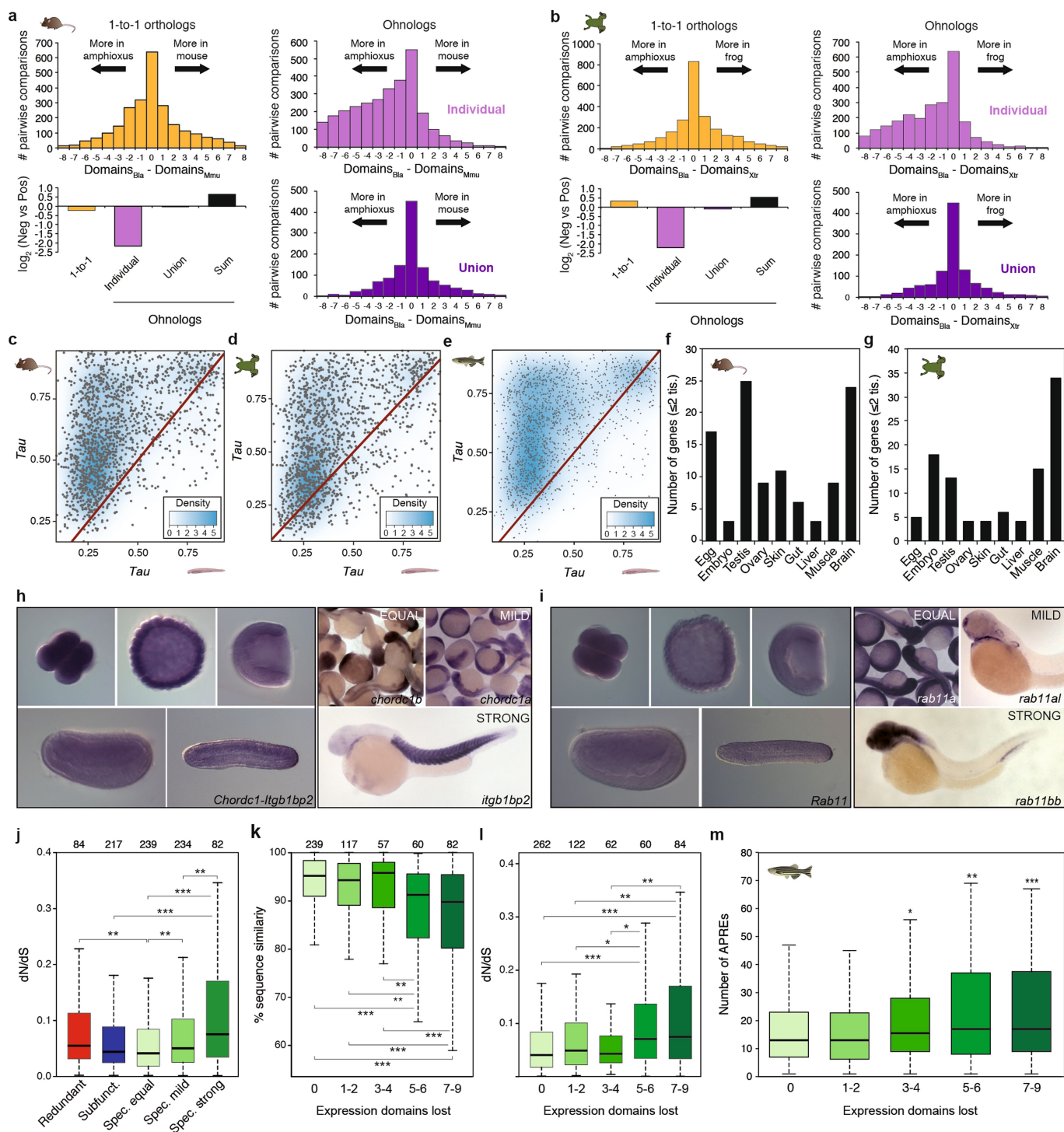
and amphioxus, 8,633. **c**, Differentially enriched GO terms among pairs of zebrafish and amphioxus Mfuzz clusters with significant orthologue overlap ( $P < 10^{-10}$  upper-tail hypergeometric test) with homochronic (48 pairs) and heterochronic (35 pairs) patterns. The GO enrichment of a group was calculated as the number of cluster pairs with significant enrichment for that given term (Supplementary Data 2, dataset 12). **d**, Top, per cent of zebrafish genes from each developmental pathway we studied, based on the temporal similarity of their corresponding Mfuzz cluster (homochronic, heterochronic or intermediate). Only genes belonging to clusters with significant orthologue overlap were analysed; the number of genes is provided in parenthesis below the pathway name. Bottom, pairwise comparisons between developmental pathway distributions.  $P$  values correspond to Bonferroni-corrected, two-sided, three-way Fisher's exact tests.



**Extended Data Fig. 9 | Higher regulatory content in vertebrate genomes.** **a**, Distribution of the number of APREs per the regulatory landscape of a gene (as determined by GREAT<sup>26</sup>), at different developmental stages or cell lines of four chordate species (amphioxus, zebrafish, medaka and mouse). Orthologous gene families are split according to the number of ohnologues that are retained per family (from 1 to 4, using mouse as a reference species for the ohnologue counts). The percentage of developmental regulatory genes (trans-dev, TD) in each category is indicated. **b**, *P* values of one-sided Mann–Whitney *U* tests against the amphioxus peak-number distribution using 100% of the minimum read coverage for different levels of down-sampling of the zebrafish and medaka samples. **c**, Distribution of the number of APREs in the GREAT region of the gene, called after down-sampling the reads of the two vertebrate samples to different fractions of the sample with the minimum effective coverage in our study (~21 reads per kbp for the 36-hpf sample in amphioxus). Asterisks correspond to the significance of the *P* values of Mann–Whitney *U* tests against the amphioxus peak-number distribution using 100% of the minimum-read coverage. The number of genes per box was as follows: amphioxus, 20,569; zebrafish,

20,053; and medaka, 15,978. **d**, As in **a**, but with gene families separated according to functional categories (housekeeping, trans-dev and others). **e**, Number of APREs per regulatory landscape determined using 4C-seq, for 58 members of 11 trans-dev families. The number of genes probed in each species is indicated on the *x* axis. **f**, Distribution of the length of the intergenic regions from the genes plotted in **a** for the indicated stages. **g**, Distributions of GREAT-region sizes (left) and number of APREs per gene (right) for a subset of 10,186 pairs of genes with matched GREAT-region size distributions (±500 bp) in amphioxus and zebrafish. **h**, Distributions of intergenic-region sizes (left) and number of APREs per gene (right) for a subset of 13,941 pairs of genes with matched intergenic-region size distributions (±500 bp) in amphioxus and zebrafish. *P* values correspond to Mann–Whitney *U* tests: \*0.05 > *P* value ≥ 0.01, \*\*0.01 > *P* value ≥ 0.001, \*\*\**P* value < 0.001. In **a** and **d**, all comparisons between each distribution of a vertebrate species and the equivalent distribution in amphioxus produced significant *P* values (*P* value < 0.001); for simplicity, in these panels asterisks are not shown. Exact *P* values and sample sizes are provided in Supplementary Data 2, dataset 8.





Extended Data Fig. 10 | See next page for caption.

**Extended Data Fig. 10 | Regulatory evolution after vertebrate WGD.**

**a, b**, For each mouse (**a**) or frog (**b**) gene, the number of positive-expression domains across nine equivalent samples is subtracted from the number of domains in which the single amphioxus orthologue is expressed. The distribution of the difference in domains between the amphioxus and the vertebrate species is plotted for 1-to-1 orthologues (2,450 and 2,484 gene pairs for mouse and frog, respectively; yellow), individual ohnologues (3,011 and 2,637 gene pairs in 1,212 and 1,094 families for mouse and frog, respectively; lilac) and the union of all vertebrate ohnologues in a family (purple). Bottom left,  $\log_2$  of the ratio between the sum of all mouse (**a**) or frog (**b**) genes with negative versus positive score for each orthology group. 'Sum' (black), binarization of family expression is performed after summing the raw expression values for all ohnologues. **c–e**, Density scattered plot of the  $\tau$  values for pairs of mouse (**c**,  $n = 1,502$ ), frog (**d**,  $n = 1,495$ ) and zebrafish (**e**,  $n = 1,498$ ) and amphioxus orthologues from multi-gene families in vertebrates. **f, g**, Number of ohnologues with strong specialization ( $\leq 2$  remaining expression domains) in mouse (**f**) or frog (**g**) expressed in each tissue or

developmental stage. **h, i**, Representative in situ hybridization assays in zebrafish embryos for different members of specialized families (right) and for the single amphioxus orthologue (left) (Chordc1 and Itgb1bp2 (**h**) and Rab11 (**i**)). Zebrafish image data for this paper were retrieved from the Zebrafish Information Network (ZFIN), University of Oregon, Eugene, OR 97403-5274; (<http://zfin.org/>, accessed May 2018) and are used with the permission of B. Thisse. Amphioxus in situ hybridization was performed once using 10 embryos per probe, all of which showed the same expression pattern. **j**, Distribution of the dN/dS ratio between human and mouse for different classes of ohnologues based on their fate after WGD. **k, l**, Distribution of the percentage of nucleotide sequence similarity (**k**) or dN/dS ratio (**l**) between human and mouse for ohnologues grouped by the number of expression domains lost. **m**, Distribution of the number of APREs within GREAT regions for zebrafish ohnologues grouped by the number of expression domains lost. *P* values in **j–m** correspond to Wilcoxon sum-rank tests. \* $0.5 > P \text{ value} \geq 0.01$ ; \*\* $0.01 > P \text{ value} \geq 0.001$ ; \*\*\* $P \text{ value} < 0.001$ .

## Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

### Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a Confirmed

- ☐ ☒ The exact sample size ( $n$ ) for each experimental group/condition, given as a discrete number and unit of measurement
- ☒ ☐ An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☐ ☒ The statistical test(s) used AND whether they are one- or two-sided  
*Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
- ☒ ☐ A description of all covariates tested
- ☐ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☐ ☒ A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☐ ☒ For null hypothesis testing, the test statistic (e.g.  $F$ ,  $t$ ,  $r$ ) with confidence intervals, effect sizes, degrees of freedom and  $P$  value noted  
*Give  $P$  values as exact values whenever suitable.*
- ☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☒ ☐ Estimates of effect sizes (e.g. Cohen's  $d$ , Pearson's  $r$ ), indicating how they were calculated
- ☐ ☒ Clearly defined error bars  
*State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on [statistics for biologists](#) may be useful.

### Software and code

Policy information about [availability of computer code](#)

#### Data collection

Images of reporter assays in zebrafish embryos were collected using cell Sens Entry 1.6 from Olympus Corporation, amphioxus ISH images with QED Capture Version 2.0.33 and amphioxus reporter assays with LAS AF (Leica Application Suite Advanced Fluorescence, version 2.7.3.). CAGE data from the FANTOM repository were obtained with CAGEr. Zebrafish CAGE-seq data were obtained through the ZebrafishDevelopmentalCAGE R package available at <http://promshift.genereg.net/CAGEr/>.

#### Data analysis

Custom code is available at: <https://gitlab.com/FunctionalAmphioxus>

The following packages and software were used:

1. FastXend, based on fastx v0.0.13.1 (<https://github.com/institut-de-genomique/fastxtend>)
2. SOAP de novo (v2.04, release 13 Jul 2012)
3. Bowtie2 (v 2.2.6)
4. Bowtie (v1.1.2)
5. MACS2 (v2.1.1.20160309)
6. peakC (v0.2)
7. Jellyfish (v2.2.0)
8. Musket (v1.0.6)
9. Flash (v1.2.1)
10. Gapcloser (v1.12)



11. Haplomerger pipeline (release 20111230)
12. Lastz (v1.02.00 release 20100112)
13. CEGMA (v2.4, rel Nov 2006)
14. EVIDENCE Modeler, EVM (r2012-06-25)
15. TransDecoder (v2.0.1)
16. AUGUSTUS (v2.5.5)
17. Exonerate (v2.2.0)
18. Trinity (v2.2.0)
19. PASA (v2.0.2)
20. Tophat2 (v2.1.1)
21. Cufflinks (v2.2.1)
22. Cuffmerge (v1.0.0)
23. CPAT software (v1.2.1)
24. InterProScan (v5.10-50)
25. TopGO (R package version 2.29.0, R version: 3.2.4)
26. RepeatScout (v1.0.5)
27. RepeatModeler (v1-0-8)
28. RepeatMasker (open-4.0.5) (<http://www.repeatmasker.org>)
29. UCSC Kent utilities (downloaded Oct-22-2013): axtChain, chainMergeSort, chainPreNet, chainNet, multiz-tba (version 2009-Jan-21)
30. PhastCons (v1.3)
31. OMA (v1.0.5)
32. RAXML (v8.2.1)
33. ETE 3 python library (v3.0.0b36)
34. MAFFT (v7.2.2.2)
35. idr (v2.0.3)
36. Kallisto (v0.42.4)
37. preprocessCore package from Bioconductor (v1.40.0)
38. Mfuzz (Program version: v2.40.0; R version: 3.2.4)
39. Weighted Gene Correlation network analysis (WGCNA) (Program version: v1.62; R version 3.3.2)
40. CAGEr (Bioconductor v1.18.0)
41. heatmaps package, Bioconductor (v.1.4.0)
42. NucleoATAC (v0.3.2)
43. Bedtools (v2.26.0)
44. deepTools (v2.4)
45. DiffBind (v2.2.3; R version: 3.3.1)
46. WALT (v1.0) (<https://github.com/smithlabcode/walt>)
47. MethylDackel (v0.3) (<https://github.com/dpryan79/MethylDackel>)
48. HMMER (v3.1b)
49. GimmeMotifs (v0.11.1)
50. Matt v.1.2.0 (<http://matt.crg.eu/>)
51. ImageJ Fiji 64
52. jupyter-notebooks (v5.1.0)
53. Photoshop CS5.1
54. Inkscape (v0.92.1 r15371)
55. Illustrator CS5.1
56. Python (v2.7)
57. Python (v3.5)
58. Python libraries: numpy (1.13.3), scipy (1.0.0), pandas (0.19.2), sklearn (0.18.1), pybedtools(0.7.10), pysam (0.12.0.1), matplotlib (2.1.0), seaborn (0.8.1)
59. GNU/Linux Command-Line Tools: zcat (1.5) , awk (4.0.2), GNU coreutils 8.22 (shuf,cat,sort,uniq,wc,head,tail)
60. Blast (v2.2.29)
61. perl (v5.14.12)
62. Pfam database (v30.0)
63. vioplot (v0.2; R version: 3.2.4)
64. alluvial (v0.1-2; R version: 3.4.2)
65. methylpy (<https://bitbucket.org/schultzmatt/methylpy.git>)
66. fastp (v0.12.5)
67. samtools (v1.6)
68. kentUtils (v302.1)
69. ggplot2 (v3.0.0; R version: 3.5.0)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

## Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Next generation sequencing data have been deposited in Gene Expression Omnibus (GEO) under the following accession numbers: GSE106372 (ChIP-seq),

GSE106428 (ATAC-seq), GSE106429 (CAGE-seq), GSE106430 (RNA-seq), GSE102144 (MethylC-seq and RRBS), and GSE115945 (4C-seq). Raw genome sequencing data and genome assembly have been submitted to European Nucleotide Archive (ENA) under the accession number PRJEB13665.

## Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/authors/policies/ReportingSummary-flat.pdf](https://www.nature.com/authors/policies/ReportingSummary-flat.pdf)

## Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical method was used to predetermine sample size. For each cross-species transcriptomic and epigenomic comparisons as many orthologous genes as possible were used. Samples sizes for each analysis in the figures are indicated in the legends, Supplementary Dataset 8 and/or Supplementary Information. For each next generation sequencing experiment, for each of the biological replicate we used as many embryos or amount of adult tissues as necessary to obtain the enough amount of RNA/DNA for library preparation and sequencing, according to previous experience in the lab and previous publications (PMIDs: 26829752, 26928226, 23624103). These numbers are indicated in the corresponding sections of the Supplementary Information.
Data exclusions	For the CAGE-seq analysis, the muscle sample was excluded, as it did not fulfill the standard quality checks. Exclusion criteria for CAGE data are not predetermined, however it is long established that CAGE data has a characteristic variation in widths (e.g. Carninci et al. 2006, Nature Genetics 38 (6): 626–35), and this, along with the very low number of reads recovered (Supplementary Dataset 1) was the rationale for exclusion, as explained in Section 7.1 of Supplementary Information.
Replication	Nearly all the findings reported in this study correspond to computational analyses of next generation sequencing data. We provide the code and guidelines to reproduce all the analyses ( <a href="https://gitlab.com/FunctionalAmphioxus">https://gitlab.com/FunctionalAmphioxus</a> ). We also perform two main types of experiments, largely for validation purposes: (i) generation of transgenic assays and (ii) in situ hybridization of specialized families. For (i), we provide the number of independent founders identified for each tested element and a description of the patterns obtained for each founder in Supplementary Table 8. For (ii), we have performed the in situ hybridization only once, using ten embryos for probe, and all of them showed the same pattern.
Randomization	We did not have experimental groups that apply here. In our study we compared either (i) different tissues and developmental stages within a species, or (ii) matched samples for different species.
Blinding	Blinding was not relevant to our study since we did not have experimental groups to compare.

## Reporting for specific materials, systems and methods

### Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Unique biological materials
<input type="checkbox"/>	<input checked="" type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input type="checkbox"/>	<input checked="" type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants

### Methods

n/a	Involved in the study
<input type="checkbox"/>	<input checked="" type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

### Antibodies

Antibodies used	-Rabbit polyclonal to Histone H3 (tri methyl K4) - ChIP Grade (#ab8580, Abcam), 1:200 -Rabbit polyclonal to Histone H3 (acetyl K27) - ChIP Grade (#ab4729, Abcam), 1:200 -Mouse monoclonal to Histone H3 (tri methyl K27) - ChIP Grade (#ab6002, Abcam), 1:200
Validation	The three primary antibodies used are all high-quality commercial antibodies against Histone H3 modifications, validated as ChIP grade by the manufacturer (Abcam): <a href="https://www.abcam.com/histone-h3-tri-methyl-k4-antibody-chip-grade-ab8580.html">https://www.abcam.com/histone-h3-tri-methyl-k4-antibody-chip-grade-ab8580.html</a> <a href="https://www.abcam.com/histone-h3-acetyl-k27-antibody-chip-grade-ab4729.html">https://www.abcam.com/histone-h3-acetyl-k27-antibody-chip-grade-ab4729.html</a> <a href="https://www.abcam.com/histone-h3-tri-methyl-k27-antibody-mabcam-6002-chip-grade-ab6002.html">https://www.abcam.com/histone-h3-tri-methyl-k27-antibody-mabcam-6002-chip-grade-ab6002.html</a>

the Histone Antibody Specificity Database (<http://www.histoneantibodies.com>, PMID : 26212453) and the Antibody Validation Database (<http://compbio.med.harvard.edu/antibodies/>). Furthermore, these antibodies have been validated by ChIP-seq in previous publications in a wide range of animal species, all of which have 100% amino acid identity in their Histone 3 sequences (PMIDs: 24642862, 22593555, 22196729).

## Animals and other organisms

Policy information about [studies involving animals](#); [ARRIVE guidelines](#) recommended for reporting animal research

Laboratory animals	For zebrafish and medaka data, embryos were obtained from crosses of adult animals of the AB and Cab strains, respectively. Gender was not determined, since we have used pools of embryos.
Wild animals	The study did not involve animals in the wild.
Field-collected samples	Most of the next-generation sequencing data generated in this study is from the amphioxus <i>Branchiostoma lanceolatum</i> , a marine, non-cephalopod invertebrate. As indicated in the Methods section, adult specimens were collected from the wild at the Racou beach near Argelès-sur-Mer, France, (latitude 42° 32' 53" N and longitude 3° 03' 27" E) with a specific permission delivered by the Prefect of Region Provence Alpes Côte d'Azur. <i>Branchiostoma lanceolatum</i> is not a protected species. Gametes are collected by heat stimulation. Embryos are collected from in vitro fertilization after induced spawning, and often batches of several dozens or hundreds of mixed embryos are used for the experiments (detailed in Methods). Adult tissues are collected from several individuals of both genders, except for male and female gonads.

## ChIP-seq

### Data deposition

- ☒ Confirm that both raw and final processed data have been deposited in a public database such as [GEO](#).
- ☒ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links  
*May remain private before publication.*

<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE106372>

Files in database submission

GSM2836695	ChIPseq amphioxus H3K27ac 15h replicate 1	Oct 31, 2020	approved	BED
GSM2836696	ChIPseq amphioxus H3K27ac 15h replicate 2	Oct 31, 2020	approved	BED
GSM2836697	ChIPseq amphioxus H3K27ac 36h replicate 1	Oct 31, 2020	approved	BED
GSM2836698	ChIPseq amphioxus H3K27ac 36h replicate 2	Oct 31, 2020	approved	BED
GSM2836699	ChIPseq amphioxus H3K27ac 8h replicate 1	Oct 31, 2020	approved	BED
GSM2836700	ChIPseq amphioxus H3K27ac 8h replicate 2	Oct 31, 2020	approved	BED
GSM2836701	ChIPseq amphioxus H3K27me3 15h replicate 1	Oct 31, 2020	approved	BED
GSM2836702	ChIPseq amphioxus H3K27me3 15h replicate 2	Oct 31, 2020	approved	BED
GSM2836703	ChIPseq amphioxus H3K27me3 36h replicate 1	Oct 31, 2020	approved	BED
GSM2836704	ChIPseq amphioxus H3K27me3 36h replicate 2	Oct 31, 2020	approved	BED
GSM2836705	ChIPseq amphioxus H3K27me3 8h replicate 1	Oct 31, 2020	approved	BED
GSM2836706	ChIPseq amphioxus H3K4me3 15h replicate 1	Oct 31, 2020	approved	BED
GSM2836707	ChIPseq amphioxus H3K4me3 15h replicate 2	Oct 31, 2020	approved	BED
GSM2836708	ChIPseq amphioxus H3K4me3 36h replicate 1	Oct 31, 2020	approved	BED
GSM2836709	ChIPseq amphioxus H3K4me3 36h replicate 2	Oct 31, 2020	approved	BED
GSM2836710	ChIPseq amphioxus H3K4me3 8h replicate 1	Oct 31, 2020	approved	BED
GSM2836711	ChIPseq amphioxus H3K4me3 8h replicate 2	Oct 31, 2020	approved	BED

Genome browser session  
(e.g. [UCSC](#))

[https://genome-asia.ucsc.edu/cgi-bin/hgTracks?db=hub\\_78274\\_BraLan2](https://genome-asia.ucsc.edu/cgi-bin/hgTracks?db=hub_78274_BraLan2)

### Methodology

Replicates

As described in Fig 1A, all samples were done in duplicates, with the exception of H3K27me3 for 8hpf, for which only one library could be constructed.

Sequencing depth

H3K27ac\_8hpf\_a 8 hpf 49 SE 18,380,227  
H3K27ac\_8hpf\_b 8 hpf 49 SE 47,674,217  
H3K27ac\_15hpf\_a 15 hpf 49 SE 11,788,559  
H3K27ac\_15hpf\_b 15 hpf 49 SE 15,294,563  
H3K27ac\_36hpf\_a 36 hpf 49 SE 42,320,156  
H3K27ac\_36hpf\_b 36 hpf 49 SE 35,867,629  
H3K27me3\_8hpf\_a 8 hpf 49 SE 11,896,658  
H3K27me3\_15hpf\_a 15 hpf 49 SE 40,188,161  
H3K27me3\_15hpf\_b 15 hpf 49 SE 39,334,807  
H3K27me3\_36hpf\_a 36 hpf 49 SE 11,969,177  
H3K27me3\_36hpf\_b 36 hpf 49 SE 12,496,630  
H3K4me3\_8hpf\_a 8 hpf 49 SE 53,565,742  
H3K4me3\_8hpf\_b 8 hpf 49 SE 23,484,131  
H3K4me3\_15hpf\_a 15 hpf 49 SE 11,831,374



H3K4me3\_15hpf\_b 15 hpf 49 SE 58,501,168  
H3K4me3\_36hpf\_a 36 hpf 49 SE 33,948,863  
H3K4me3\_36hpf\_b 36 hpf 49 SE 32,457,194

#### Antibodies

-Rabbit polyclonal to Histone H3 (tri methyl K4) - ChIP Grade (#ab8580, Abcam)  
-Rabbit polyclonal to Histone H3 (acetyl K27) - ChIP Grade (#ab4729, Abcam)  
-Mouse monoclonal to Histone H3 (tri methyl K27) - ChIP Grade (#ab6002, Abcam)

#### Peak calling parameters

Reads were mapped against the amphioxus reference genome using Bowtie, and peaks were called using the MACS2 software with default parameters.

#### Data quality

Chip-seq peaks were only used to overlap with the ATAC-seq peaks in multiple cross-validation analyses. Since we used MACS2 by default, all peaks were below FDR 5%.

#### Software

Reads were mapped against the amphioxus reference genome using Bowtie, and peaks were called using the MACS2 software with default parameters. The overlap between ATAC-seq and ChIP-seq peak was calculated using Bedtools.